4

The Economics of Standards

OLE HANSETH

Weill and Broadbent (1998: 7) see an organization's collection of hardware, software, devices, data, and IT-related personnel as its IT infrastructure. They describe this as an IT portfolio, which should be regarded as any other investment portfolio. This investment-portfolio metaphor can certainly be useful for understanding some aspects of IT infrastructures. But it is a metaphor that can also be very misleading (see Chapter 2).

Investment portfolios are usually very flexible and easy to change, manage, and control. Elements of such portfolios may be sold at almost any time, and individual elements might be sold or bought independently (although portfolios should be balanced to minimize risks, and so on). Infrastructures are different. The individual elements are very interdependent, and their size and complexity may make them extremely difficult to control and manage. In this chapter we will present basic findings and concepts from economic studies of infrastructures and standards. These studies are focused on public standards and infrastructures. But, as we will argue below, the results are equally valid for corporate IT infrastructures.

Here we see another deficiency of the management literature. On the one hand, it argues for the importance of *corporate* information infrastructure by making reference to the vital role played by *public* infrastructures at national and international levels. But, on the other hand, it stays clear of all the intricacies and dilemmas that the development and management of such public infrastructures imply, and that the economic studies reported below try to discuss.

Infrastructure and Standards

We will here identify a few key characteristics of infrastructures. These are aspects that we see as essentials of traditional, public infrastructures, and at the same time characteristics that make IT infrastructures different from traditional information systems.

In Webster's *Dictionary* infrastructure is defined as 'a sub-structure or underlying foundation; esp., the basic installations and facilities on which the continuance and growth of a community, state, etc. depend as roads, schools, power plants, transportation and communication systems, etc.' (Guralnik 1970).

One aspect of infrastructures that we can extract out of this definition is that they have a supporting or *enabling* function. An infrastructure is designed to support a wide range of activities; it is not especially tailored to one. It is enabling in the sense

that it is a technology intended to open up a field of new activities, not just to improve or automate something that already exists. This is opposed to being designed especially to support one way of working within a specific application field. The enabling feature of infrastructures plays an important role in policy documents, such as the Clinton–Gore and Bangemann *et al.* (1994) reports on national and European information infrastructures respectively.

Although infrastructures are considered to be enabling, they are, of course, designed to provide certain supporting functions. The interdependencies between the specific *design* of infrastructural services and their *use* is often overlooked. Joerges (1988) has pointed to the fact that infrastructures are subject to 'deep ecological penetration'. This penetration takes place very gradually and over a long time, which to a large extent may make us blind to its existence as well as its implications.

A second aspect we can extract from Webster's definition is the fact that an infrastructure is *shared* by a larger community (or collection of users and user groups). An infrastructure is shared by the members of a community in the sense that it is the same single object used by all of them (although it may appear differently). In this way infrastructures should be seen as irreducible; they cannot be split into separate parts to be used independently by different groups. An e-mail infrastructure is one such shared irreducible unit. An example of the opposite is the various installations of a word processor that may be used completely independently of each other. However, an infrastructure may, of course, be broken down into separate units for analytical or design purposes.

The different elements of an infrastructure are integrated through *standardized interfaces*. Often it is argued that such standards are important, because the alternative—bilateral—arrangements are all too expensive. This is certainly true. However, standards are not only economically important but also a necessary constituting element. If an 'infrastructure' is built on the bases of bilateral arrangements only, this is not a real infrastructure, but just a collection of independent connections.

The shared and enabling aspects of infrastructures have made the concept increasingly popular over recent years. Just as in the case of information infrastructures, the role of infrastructures is believed to be important, as its enabling character points to what may be kept as a stable basis in an increasingly more complex and dynamic world. Håkon With-Andersen (1997) documents how the Norwegian shipbuilding sector has stayed competitive through major changes from sailing ships, through steamboats and tankers, to offshore oil drilling supply boats, because of the crucial role of an infrastructure of competence centres and supporting institutions such as shipbuilding research centres, a ship classification company, and specialized financial institutions. In the same way, Ed. Steinmueller (1996) illustrates how the shared character of infrastructures is used to help understand the growing importance of knowledge as public good and infrastructure in societies where innovation and technological development are crucial for the economy.

The third aspect of infrastructures we will emphasize is their *openness*. They are open in the sense that there are no limits to the number of users, stakeholders,

and vendors involved, nodes in the network and other technological components, application areas, network operators, and so on. This defining characteristic does not necessarily imply the extreme position that absolutely everything is included in every information infrastructure. However, it does imply that one cannot draw a strict border saying that there is one infrastructure for what is on one side of the border and others for the other side and that these infrastructures are independent.

Whatever the numbers of an infrastructure's user groups, application areas, designers and manufacturers, network operators, service providers, and so on, there will always be someone or something outside that should be involved or to which the infrastructure should be connected.

Unlimited numbers of users, developers, stakeholders, components, and use areas imply:

- several activities with varying relations over time;
- varying constellations and alliances;
- changing conditions for development;
- changing requirements.

In sum, all this leads to heterogeneity.

In the Clinton–Gore report, the envisioned NII (or 'electronic superhighway') is meant to include more than just the physical facilities used to transmit, store, process, and display voice, data, and images. It is considered to encompass the following.

- A wide and ever-expanding range of *equipment*, including cameras, scanners, keyboards, telephones, fax machines, computers, switches, compact disks, video and audio tape, cable, wire, satellites, optical fibre transmission lines, microwave nets, switches, televisions, monitors, printers, and much more.
- The *information* itself, which may be in the form of video programming, scientific or business databases, images, sound recordings, library archives, other media, and so on
- *Applications* and *software* that allow users to access, manipulate, organize, and digest the proliferating mass of information that the NII's facilities will put at their fingertips.
- The network *standards* and transmission codes that facilitate interconnection and interoperation between networks.
- The *people* who create the information, develop applications and services, construct the facilities, and train others to tap its potential.

The report says that every component of the information infrastructure must be developed and integrated if the USA is to capture the promise of the Information Age.

This definition also sees infrastructures as enabling, shared, and open. Further, it points to some other crucial features we now will turn to. Infrastructures are *heterogeneous* phenomena. They are so in at least two ways.

First, information infrastructures are more than 'pure' technology; they are rather *socio-technical networks*. Infrastructures are heterogeneous concerning the qualities of their constituencies. They include technological components, humans, organizations, institutions, and so on. This fact is most clearly expressed in the last point in the list above. It is true for information technologies in general, as they will not work without support staff. Nor will an information system work if the users are not using it properly. For instance, flight-booking systems do not work unless all booked seats are registered in the systems.

Secondly, infrastructures are connected and interrelated, constituting *ecologies of infrastructures*. One infrastructure is composed of ecologies of (sub)infrastructures by

- building one infrastructure as a layer on top of another;
- linking logical related networks;
- integrating independent components, making them interdependent.

Infrastructures are layered upon each other just as software components are layered upon each other in all kinds of information systems. This is an important aspect of infrastructures, but one that is easily grasped as it is so well known. An example is the World Wide Web as a global infrastructure built on top of the Internet's global TCP/IP infrastructure.

Infrastructures are also heterogeneous in the sense that the same logical function might be implemented in several different ways. Larger infrastructures will often be developed by interconnecting two existing different ones, as has typically happened when networks such as America Online and Prodigy have been connected to the Internet through gateways. In principle the same happens when one standardized part (protocol) of an infrastructure is being replaced over time by a new one. In such transition periods, an infrastructure will consist of two interconnected networks running different versions. A paradigm example of this phenomenon is the transition of the Internet from IPv4 to IPv6 (Hanseth *et al.* 1996; Monteiro 1998).

Infrastructures are also heterogeneous in the sense that larger components or infrastructures are built based on existing smaller, independent components. When these components are brought together into a larger unit, they become interdependent. When one of them is changed, for whatever reason, the others will often need to be changed as well. Examples of this phenomenon are the various formats for representing text, video, sound, image, and graphical representations that are brought together and put into MIME to enable transfer of multimedia information on the Web/Internet. At the time of writing, the Internet integrates a vast range of standardized formats and protocols, in total more that 200 standards (RFC 1994). The latest version of IP, IPv6, contains only modest changes to the former. However, it requires fifty-one other Internet standards to be changed.

Building large infrastructures takes *time*. All elements are connected. As time passes, new requirements appear to which the infrastructure has to adapt. The whole infrastructure cannot be changed instantly—the new has to be connected to the old. The new version must be designed in such a way as to make the old and

the new interlinked and 'inter-operable' in one way or another. In this way the old—the installed base—heavily influences how the new can be designed. This leads us to the last aspect of infrastructures that we want to point out: they develop through extending and improving the *installed base*.

The focus on infrastructure as an 'installed base' implies that infrastructures are always considered as existing already; they are NEVER developed from scratch. When 'designing' a 'new' infrastructure, it will always be integrated into or replace part of an existing one. This has been the case in the building of all transport infrastructures. Every single road—even the first one, if it makes sense to speak of such a thing—has been built in this way; when air-traffic infrastructures have been developed, they have been tightly interwoven with road and railway networks—for these are needed for travel between airports and travellers' destinations. Moreover, air-traffic infrastructures can be used for only one part of a journey, and isolated air-traffic infrastructures, without the support of other infrastructures, would be useless.

The notion of an installed base does to a large extent include all the aspects of infrastructure mentioned above: an *infrastructure is an evolving shared, open, and heterogeneous installed base.*

Public versus Corporate Infrastructures

The aspects of infrastructures pointed out above are derived from how we see traditional public infrastructures. Are public and corporate infrastructures essentially of the same nature, so that these aspects are also the most important ones for corporate information infrastructures? We believe so.

Of the aspects mentioned, openness is the most crucial one in this respect. In general we can say that (public) infrastructures are shared by open communities, while (information) systems (in particular as presented in IS development methodology and strategy textbooks) are used by closed organizations. This is important with regard to the size and complexity of infrastructures, but above all with regard to their governance—that is, what makes the control of the design and use of infrastructures different from that of information systems.

Infrastructures are primarily designed through the standardization of interfaces and protocols and through the diffusion of the various standardized components. The actual infrastructure can be said to be designed through the latter processes. The Internet, for instance, is designed (as a network) as the individual users and Internet Service Providers install Internet software on their computers and link them to the existing network. Both the standardization and infrastructure building activities are carried out by a huge number of independent actors. Some institutions, like standardization bodies, are set up in order to try to coordinate such processes. But these institutions have hardly any authority or power to enforce any kind of behaviour on the individual actors. This is the opposite of the picture usually drawn of organizations. The hierarchical structure was invented to control and coordinate processes. Any manager has the authority to make decisions and

instruct his or her subordinates what to do. But this is an ideal picture and not the reality.

Modern global corporations are more like an open community than a closed organization—at least as far as their information infrastructures are concerned. The size, variety, and complexity of global organizations such as those presented in this volume make them difficult to manage as one coherent unit. This is possible only for a few issues to which top management dedicate their resources. The character of global corporations as a collection of a large number of independent units is also a result of their dynamics and modern management models emphasizing fast decision-making, flexibility, and accordingly local autonomy (see Chapter 3). Such models become more important as the company grows and its level of competence and knowledge intensity increase. Further, the technological development that all IT activities in a corporation depend on is external. It is controlled externally and not by the corporation. In addition, the companies are becoming less manageable as they become more deeply embedded in their environment—for instance, through closer collaboration with suppliers, customers, and other strategic partners. The last point is of particular relevance for the governance of infrastructures. This implies that the information infrastructures of modern organizations are to a large extent shared with collaborating organizations; they are becoming public infrastructures. All these aspects are, as shown in the previous chapter, becoming more important, as organizations are becoming increasingly more modern and global.

A crucial aspect of infrastructure development is the design and diffusion of *standards*. This is also valid for corporate IT infrastructures. Weill and Broadbent (1998: 266) say that, to succeed in the establishment of IT infrastructures, a corporation should 'enforce IT architecture and standards'. This statement is presented without any explanation, justification, or arguments, which makes one believe that the authors see this as an obvious truth and a plain and simple guideline to follow in practice. We believe that the opposite is the case. Below we will enquire into the nature of standardization definition and implementation processes by presenting the basic lessons learned by economists who have studied standardization processes.

Economics of Standards

The main concepts within the 'economics of standards' that should attract our attnetion are: increasing returns and positive feedback, network externalities, path dependency, and installed base.

Increasing returns and positive feedback

Increasing returns mean that, the more a particular product is produced, sold, or used, the more valuable or profitable it becomes. Infrastructure standards are paradigm examples of products having this characteristic. The development and

diffusion of 'infrastructural' technologies are determined by 'the overriding importance of standards and the installed base compared to conventional strategies concentrating on programme quality and other promotional efforts' (Grindley 1995: 7).

A communication standard's value is to a large extent determined by the number of users—that is, the number of users you can communicate with if you adopt the standard. This phenomenon is illustrated by well-known examples such as Microsoft Windows and the rapid diffusion of the Internet in recent years. Earlier examples are the sustainability of FORTRAN and COBOL far beyond the time when they had become technologically outdated.

The basic mechanism is that the large installed base attracts complementary products and makes the standard cumulative more attractive. A larger base with more complementary products also increases the credibility of the standard. Together these make a standard more attractive to new users. This brings in more adoptions, which further increases the size of the installed base, and so on, as illustrated by Fig. 4.1 (Grindley 1995: 27).

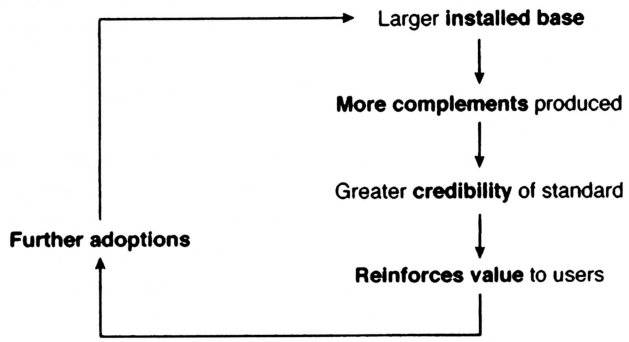

FIG. 4.1. Standards reinforcements mechanism
Source: Grindley (1995).

The phenomenon of positive feedback has been focused and theorized over recent years in studies of standards. The concept is opposed to basic assumptions of classical economics. Classical economics is centred on the notion of negative feedback or decreasing returns. These processes hold for economies based on natural resources such as agriculture and mining, which are subject to diminishing returns caused by limited amounts of fertile land or high-quality ore deposits. In such cases, the most fertile and easily available land and the most easily available oilfields are utilized first. As more agricultural products or oil are produced, so the land that is more difficult to access, or the more marginal resources, have to be used, and production becomes more expensive (Arthur 1994).

Increasing returns stem from more than one source. The most widely known one is the fact that larger firms tend to have smaller unit costs. This source of positive feedback is known as scale economies—or, more precisely, economies of scale in production (Shapiro and Varian 1999).

Economies of scale, again, have several causes. Large-scale production machinery produces units at lower cost. But, in some sectors, the costs of developing new products are large, while making copies of the product is cheap. This phenomenon is particularly important in high-tech and research-intensive industries, and in the software and information sector more than in any other, as the unit costs of making software and information products (like a digital encyclopedia) are close to zero, while the product development costs can be extremely high.

Further, there is a strong connection between increasing-returns mechanisms and *learning* processes. Increased production brings additional benefits: producing more units means gaining more experience in the manufacturing process, achieving greater understanding of how to produce additional products even more cheaply. Moreover, *experience* gained with one product or technology can make it easier to produce new products incorporating similar or related technologies.

What is most focused in the economics of standards, however, is rather demand-side increasing returns and demand-side economies of scale. Demand-side economies of scale are partly created by learning effects. The transmission of information based on experience may serve as a reinforcement for early leading positions and so act in a manner parallel to more standard forms of increasing returns. A similar phenomenon is individual learning, where again success reinforces some courses of action and inhibits others, thereby causing the first to be sampled more intensively, and so forth (Arrow 1994).

Where learning takes place, beliefs can become self-reinforcing (Arthur 1994), and, accordingly, the product that is expected to become the standard will become the standard. Self-fulfilling expectations are one manifestation of positive-feedback economics and bandwagon effects. If customers expect your product to become popular, a bandwagon will form, a virtuous cycle will begin, and customers' expectations will prove correct. Success and failure are driven as much by expectations and luck as by the underlying value of the product (Shapiro and Varian 1999).

Although the concept of increasing returns has attracted the focus of many economists, it is not a new one. It has had a long but uneasy presence in economic analysis. The opening chapters of Adam Smith's Wealth of Nations put great emphasis on increasing returns to explain both specialization and economic growth. But this tradition acts like an underground river, springing to the surface every few decades (Arrow 1994). Increasing returns have been identified within traditional economies too. Manufacturing, for instance, enjoyed increasing returns because large plants allowed improved organization and created economies of scale. But the theory of 'increasing returns' is argued in particular to be the appropriate theory for understanding modern high-technology economies (Arthur 1994) and the modern emerging information economy (Shapiro and Varian 1999). For instance, Arthur (1994) argues that the economic circumstances under which a new, superior technology might replace an old inferior one, and how long this time

might take, is well explained by the theory. This 'competing technologies problem' is one of increasing returns par excellence.

According to Arthur (1994), positive feedback mechanisms are now central to modern theorizing in international trade theory, growth theory, the economics of technology, industrial organization, macroeconomics, regional economics, economic development, and political economy. Further, the part of the economy that is knowledge-based is largely subject to increasing returns. Products such as computers, pharmaceuticals, aircraft, software, and so on are complicated to design and to manufacture. They require large initial investments in research and development and tooling but, once sales begin, incremental production is relatively cheap (Arthur 1994).

Shapiro and Varian (1999) see positive feedback as the central element in the information economy, defining information as anything that may be digitized. An information good involves high fixed costs but low marginal costs. The cost of producing the first copy of an information good may be substantial, but the cost of producing (or reproducing) additional copies is negligible. They argue that the key concept in the network economy is positive feedback. The case of Microsoft is a good illustration of this. Demand-side economies of scale are the norm in information industries.

Network effects, network externalities

Whether real or virtual, networks have a fundamental economic characteristic: the value of connecting to a network depends on the number of other people already connected to it. This fundamental value proposition goes under many names: network effects, network externalities, demand-side economies of scale. They all refer to essentially the same point: other things being equal, it is better to be connected to a bigger network than a smaller one (Shapiro and Varian 1999.)

Externalities arise when one market participant affects others without compensation being paid. In general, network externalities may cause negative as well as positive effects. The classic example of negative externalities is pollution: my sewage ruins your swimming or drinking water. Positive externalities give rise to positive feedback (Shapiro and Varian 1999).

Path dependency

Network externalities and positive feedback give rise to a number of more specific effects. One such is *path dependence*. Path dependence means that past events will have large impacts on future development, and in principle irrelevant events may turn out to have tremendous effects (David 1986). For instance, a standard that builds up an installed base ahead of its competitors becomes cumulatively more attractive, making the choice of standards 'path dependent' and highly influenced by a small advantage gained in the early stages (Grindley 1995: 2). The classical and widely known example illustrating this phenomenon is the development and evolution of keyboard layouts, leading to the development and *de facto* standardization of QWERTY (David 1986).

We can distinguish between two forms of path dependence.

- Early advantage in terms of numbers of users leads to victory.
- Early decisions concerning the design of the technology will influence future design decisions.

The first one has already been mentioned above. When two technologies of a kind where standards are important—such as communication protocols or operating systems—are competing, the one getting an early lead in terms of number of users becomes more valuable for the users. This may attract more users to this technology and it may win the competition and become a *de facto* standard. The establishment of Microsoft Windows as the standard operating system for PCs followed this pattern. The same pattern was also followed by the Internet protocols during the period they were competing with OSI protocols.

The second form of path dependence concerns the technical design of a technology. When, for instance, a technology is established as a standard, new versions of the technology must be designed in a way that is compatible (in one way or another) with the existing installed base. This implies that design decisions made early in the history of a technology will often live with the technology as long as it exists. Typical examples of this are various technologies struggling with the backward compatibility problem. Well-known examples in this respect are the different generations of Intel's microprocessors, where all later versions are compatible with the 8086 processor, which was introduced into the market around 1982.

Early decisions about the design of the Internet technology, for instance, have had a considerable impact on the design of new solutions both to improve existing services and to add new ones to the Internet. For instance, the design of the TCP/IP protocol constrains how improved solutions concerning real-time multimedia transfer can be designed and how security and accounting services can be added to the current Internet.

Lock-in: switching costs and coordination problems

Increasing return may lead to yet another effect: *lock-in*. Lock-in means that, when a technology has been adopted, it will be very hard or impossible to develop competing technologies. 'Once random economic events select a particular path, the choice becomes locked-in regardless of the advantages of alternatives' (Arthur 1990). In general, lock-in arises whenever users invest in multiple complementary and durable assets specific to a particular technology. We can identify different types of lock-in: contractual commitments, durable purchases, brand-specific training, information and databases, specialized suppliers, search costs, and loyalty programmes (Shapiro and Varian 1999). We can also say that lock-ins are caused by the huge switching costs or by coordination problems (or a combination of these) that would be incurred when switching from one standardized technology to another.

Switching costs and lock-ins are ubiquitous in information systems, and managing these costs is very tricky both for sellers and buyers. For most of the history of computers, customers have been in a position where they could not avoid buying (more or less) all their equipment and software from the same vendor. The switching costs of changing computer systems could have been astronomical—and certainly so high that no organization did. To change from one manufacturer (standard) to another would imply changing all equipment and applications at the same time. This would be very expensive—far beyond what anybody could afford. But it would also be an enormous waste of resources, because the investments made have differing economic lifetimes, so there is no easy time to start using a new, incompatible system. As a result, buyers face switching costs, which effectively lock them into their current system or brand (Shapiro and Varian 1999).

Switching costs also go beyond the amount of money an organization has to pay to acquire a new technology and install it. Since many software systems are mission critical, the risks in using a new vendor, especially an unproven one, are substantial. Switching costs for customers include the risk of a severe disruption in operations.

Lock-in is not only created by hardware and software. Information itself—its structures in databases as well as the semantics of the individual data elements—is linked together into huge and complex networks that create lock-ins. One of the distinct features of information-based lock-in is that it proves to be so durable: equipment wears out, reducing switching costs, but specialized databases live on and grow, increasing lock-in over time (Shapiro and Varian 1999).

The examples of lock-ins and switching costs mentioned so far are all related to infrastructures that are seen as local to one organization. As infrastructures and standards are shared across organizations, lock-in problems become even more challenging.

Network externalities make it virtually impossible for a small network to thrive. But every network has to start from scratch. The challenge to companies introducing new but incompatible technology into the market is to build a network size that overcomes the collective switching costs—that is, the combined switching costs of all users. In many information industries, collective switching costs are the biggest single force working in favour of incumbents. Worse yet for would-be entrants and innovators, switching costs work in a non-linear way: convincing ten people connected in a network to switch to your technology is more than ten times as hard as getting one customer to switch. But you need all ten, or most of them: no one will want to be the first to give up the network externalities and risk being stranded. Precisely because various users find it so difficult to coordinate a switch to an incompatible technology, control over a large installed base of users can be the greatest asset you can have.

But lock-in is more than cost. As the community using the same technology or standard grows, switching to a new technology or standard becomes an increasingly larger *coordination* challenge. The lock-in represented by QWERTY, for instance, is most of all a coordination issue. It is shown that the individual costs

of switching are marginal (David 1986), but, as long as we expect others to stick to the standard, it is best that we do so ourselves as well. There are too many users (everybody using a typewriter or PC/computer). It is impossible to bring them together so that they could agree on a new standard and commit themselves to switch.

Many lock-in situations are of such a character that to get out of them requires both huge switching costs and coordination tasks. A typical example of such a lock-in situation is again Microsoft Windows. It is hard to imagine, at the time of writing, that any operating system can compete with Windows in the PC market, however fantastic it might be. The Linux operating systems might possibly be a competitor. But this system is developed under quite extraordinary conditions—by a huge group of enthusiastic individuals who are working without getting paid. The Internet is another example. (For more on this aspect, see Hanseth *et al.* 1996; Monteiro 1998.) It has been widely acknowledged for a long time that the TCP/IP protocol is becoming outdated. At the end of the 1990s its address space was running out, and it lacked appropriate support for wireless and mobile networks, real-time multimedia, the accounting that was required by commercial service providers, security, and so on. For these reasons, the development of a new protocol had already been launched in 1990. As the design work proceeded, the fact that the new version had to be compatible with the existing one emerged as the single most important requirement. As a consequence, other issues, except increased address space, were given up. A new protocol has been defined, but it remains to be seen whether it will be adopted by users of the Internet. The Internet example also illustrates that getting out of the lock-in trap also involves a third challenge. To avoid a lock-in, the new technological solutions must support the transition from the old to the new.

We have so far illustrated the coordination problems related to public standards and infrastructures. But there may also be huge coordination problems related to the change of corporate infrastructures. The hierarchical structure of organizations is developed for making decisions and coordination processes across the organization. Accordingly, this hierarchical structure should take care of the coordination required for changing IT infrastructures. But this will not always be possible. There will often be too many complex issues involved, too many actors and units, and the infrastructure may also be embedded in local contexts (see Chapter 5). Indeed, corporate infrastructures may be locked-in in just the same way as are public ones.

Inefficiency

The last consequence of positive feedback we mention is what is called *possible inefficiency*. This means that the best solution will not necessarily win. An illustrative and well-known example of this phenomenon is the competition between the Microsoft Windows operating system and Macintosh. Macintosh was widely held to be the best technology—in particular from a user point of view—but Windows won because it had early succeeded in building a large installed base.

Strategies

David (1987) points out three strategy dilemmas that one would usually face when developing networking technologies and that are caused by the phoneomena of network externalities and increasing returns.

- *Narrow policy window*. There may be only brief and uncertain 'windows in time' during which effective public policy interventions can be made at moderate resource costs.
- *Blind giants*. Governmental agencies are likely to have the greatest power to influence the future trajectories of network technologies just when the suitable informational basis on which to make socially optimal choices among alternatives is most lacking. The actors in question, then, resemble 'blind giants'—whose vision we would wish to improve before their power dissipates. Corporate headquarters will typically be in the same position when defining corporate standards.
- *Angry orphans*. Some groups of users will be left 'orphaned'; they will have sunk investments in systems whose maintenance and further elaboration are going to be discontinued. Encouraging the development of gateway devices linking otherwise incompatible systems can help to minimize the static economic losses incurred by orphans.

There are, in principle, two strategies to choose between to get out of a lock-in and to help avoid the dilemmas: an evolution strategy of backward compatibility or a revolution strategy of compelling performance. These strategies reflect an underlying tension when the forces of innovation meet up with network externalities: is it better to wipe the slate clean and come up with the best product possible (revolution) or to give up some performance to ensure compatibility and thus ease consumer adoption (evolution) (Shapiro and Varian 1999)?

The evolution strategy, which offers users an easy migration path, centres on reducing switching costs so that users can try your new technology gradually. In virtual networks, the evolution strategy of offering users a migration path requires an ability to achieve compatibility with existing products. In real networks, the evolution strategy requires physical interconnection to existing networks. In either case, interfaces are critical. The key to the evolution strategy is to build a new network by linking it first to the old one. The technical obstacles faced have to do with the need to develop a technology that is at the same time compatible with, and yet superior to, existing products.

The revolution strategy is inherently risky. It cannot work on a small scale and usually requires powerful allies. Worse yet, it is devilishly difficult to tell early on whether your technology will take off or crash and burn. Even successful technologies start off slowly and accelerate from there.

Radical changes are often advocated—for instance, within the business process re-engineering (BPR) literature. Empirically, however, such radical changes of larger networks are rather rare. Hughes (1987) concluded that large networks

change only in the chaos of dramatic crises (such as the oil crises in the early 1970s) or in the case of some external shock.

Gateways

We can distinguish between two variants of the evolutionary strategy—slow evolution based on *backward compatibility* and fast evolution based on *gateways* linking the new and the old networks.

The term 'gateway' has a strong connotation. It has traditionally been used in a technical context to denote an artefact that is able to translate back and forth between two different communication networks. A gateway in this sense is also called a 'converter' and operates by inputting data in one format and converting them to another. In this way a gateway may translate between two, different communication protocols that would otherwise be incompatible, as a protocol converter accepts messages from either protocol, interprets them, and delivers appropriate messages to the other protocol.

A well-known and important example of gateways, which is also analysed in the economics of standards literature, is the alternating/direct current (AC/DC) adapter (David and Bunn 1988; Hughes 1983). At the beginning of the twentieth century, it was still an open and controversial issue whether electricity supply should be based on AC or DC. The two alternatives were incompatible and a 'battle of systems' unfolded. As a user of electrical lighting, you would have had to choose between the two. There were strong proponents and interests behind both. Both had their distinct technical virtues. AC was more cost effective for long-distance transportation (because the voltage level could be higher) whereas a DC-based electrical motor preceded the AC-based one by many years. As described by Hughes (1983) and emphasized by David and Bunn (1988), the introduction of the converter made it possible to couple the two networks. It accordingly became feasible to combine the two networks and hence draw upon their respective virtues.

Other scholars have developed notions related to this notion of a gateway. Star and Griesemer's (1989) concept of *boundary objects* may also be seen as a gateway enabling communication between different communities of practices. The same is the case for Cussins' (1998) objectification strategies. These strategies may be seen as constituting different networks, each of them being connected to the networks built by the different practices through gateways translating the relevant information according to the needs of the 'objectification networks'.

Gateways fill important roles in a number of situations during all phases of an information infrastructure development. The key effect of traditional converters is that they sidestep—either by postponing or by altogether avoiding—a confrontation. The AC/DC adapter is a classic example. The adapter bought time so that the battle between AC and DC could be postponed. Hence, the adapter avoided a premature decision. Instead, the two alternatives were able to coexist and the decision to be delayed until more experience had been acquired.

Sidestepping a confrontation is particularly important during the early phases of

an infrastructure development as there is still a considerable amount of uncertainty about how the infrastructure will evolve. And this uncertainty cannot be settled up front; it has to unfold gradually. Gateways may prevent those in the position of making decisions from acting like 'blind giants'.

But it is not only during the early phases that sidestepping confrontation is vital. It is also important in a situation where there are already a number of alternatives, none of which is strong enough to 'conquer' the others. In the case of e-mail systems, for instance, many different proprietary systems and protocols were developed before the Internet or other standards were available. On this basis, it has been considered more convenient to develop the different protocols separately and link the networks together through gateways.

A more neglected role of gateways is the way they support modularization. The modularization of an information infrastructure is intimately linked to its heterogeneous character. The impossibility of developing an information infrastructure monolithically forces a more patchlike and dynamic approach. In terms of actual design, this entails decomposition and modularization. The role of a gateway, then, is to encourage this required decomposition by decoupling the efforts to develop the different elements of the infrastructure and coupling them tightly only at the end. This allows a maximum of independence and autonomy.

Modularization, primarily through black-boxing and interface specification, is, of course, an old and acknowledged design virtue for all kinds of information systems, including information infrastructures. But the modularization of an information infrastructure supported by gateways has another, essential driving force that is less obvious. As the development is more likely to take ten years than one, the contents are bound to evolve or 'drift'. As a result, previously unrelated features and functions are brought together and need to be aligned. The coupling of two (or more) of these require a highly contingent, techno-economical process, a process that is difficult to design and cater for. Cable TV and the telephone have a long-standing history of distinctly different networks. They were conceived of, designed, and appropriated in quite distinct ways. Only as a result of technological development and legislative deregulation has it become reasonable to link them. This has given rise to an *ecology of networks* that may be linked together later by gateways.

5

Actor-Network Theory and Information Infrastructure

ERIC MONTEIRO

The study of the economics of infrastructure has already begun to show how the development, introduction, and use of an information infrastructure are an involved socio-technical process of *negotiation*. The open-ended character of this process—the stumbling, the compromises, the way non-technical interests get dressed up in technical disguise—calls for an analytic vehicle that helps tease out interesting and relevant issues related to the 'management' of such processes.

This chapter introduces, outlines, and illustrates one such vehicle—namely, actor-network theory (ANT). We introduce ANT by briefly positioning it within the broader landscape of conceptualizations of technology and society. This exercise is intended to be neither comprehensive nor systematic. It is aimed at spelling out those underlying aspects of an information infrastructure towards which ANT makes us sensitive.

First and foremost, ANT, especially in the minimalistic version outlined here, offers an illuminating vocabulary to describe information infrastructure. It provides a language to describe how, where, and to what extent technology influences human behaviour. This is valuable when identifying the influence of seemingly grey and anonymous technical components such as standards or systems modules that are already installed. In particular, it allows ANT to zoom in and out of a situation as required.

This implies that the granularity (that is, the scope, depth, and level of detail) of the analysis is flexible. Sometimes a comprehensive set of interconnected modules and systems is collapsed into one node; sometimes the focus is on the relative contribution of each of the modules; sometimes a detailed analysis is needed of the design of one specific module. This kind of flexibility is indispensable in any analysis of information infrastructure.

The reason for outlining ANT in relation to the development and establishment of information infrastructure is the need critically to assess the descriptions of this issue provided by traditional management literature. This literature—as discussed in Chapter 2—is dominated by top-down, rational, decision making.

There are, of course, alternative perspectives on strategic information systems in general and information infrastructures in particular. For example, there exists an