Understanding Information Infrastructure

Ole Hanseth and Eric Monteiro

Manuscript 27. Aug., 1998

CHAPTER 1 *Introduction* ***1***

The emergence of information infrastructures **1**
Infrastructures are different **4**
New trends, new research **6**
Our focus, approach and contribution **7**
Brief outline of the book **9**
Earlier papers **14**

CHAPTER 2 *Cases: The Internet and Norwegian Health Care* ***17***

Internet **17**
International Standardization Organization (ISO) **25**
EDI and EDIFACT **26**
Health care information infrastructures **27**
EDI Infrastructure in the Norwegian health care sector **28**
Methodological issues **33**

CHAPTER 3 *Defining information infrastructures* ***37***

Introduction **37**
Roots **39**
Defining II - identifying the aspects **40**
Sliding boundaries **50**
State of the art **51**

CHAPTER 4 *Standards and standardization processes* ***55***

Introduction **55**

The economy of scale argument **56**
Types of standards **57**
Internet standards **58**
Standards for health care information infrastructures **63**
Standardisation processes and strategies **66**
OSI **66**
EDIFACT **68**
Internet **69**
Standardization of health care IIs **72**
Standards engineering **73**
State of the art in research into standardisation **75**

CHAPTER 5 *Openness and flexibility* ***77***

Introduction **77**
Openness **78**
The wider environment - a more rapidly changing world **81**
Implications: the need for flexible infrastructures **83**
Enabling flexibility to change **86**
Hampering flexibility to change **87**
The interdependencies between standardization and flexibility **88**

CHAPTER 6 *Socio-technical webs and actor network theory* ***91***

Introduction **91**
Technology and society: a brief outline **91**
The duality of technology **93**
Actor network theory - a minimalistic vocabulary **96**

Actor networks meet structuration theory **101**

CHAPTER 7 *Inscribing behaviour* ***105***

Introduction **105**
Translations and design **106**
Inscriptions and materials **110**
Accumulating the strength of an inscription **116**
Moving on **124**

CHAPTER 8 *Dreaming about the universal* ***127***

Introduction **127**
What is universalism? **128**
Universalism at large: a few illustrations **129**
The origin of universalism **136**
How to create universals **139**
Universals in practise **141**
If universalism is dead, what then? **146**

CHAPTER 9 *Installed base cultivation* ***147***

Introduction **147**
Installed base **147**
The irreversibility of the installed base **148**
Information infrastructure as actor **149**
Effects **151**
Some examples **153**
Strategy **154**
From design to cultivation **156**
Cultivating the installed base **160**

CHAPTER 10 *Changing infrastructures: The case of IPv6* ***165***

Introduction **165**
The IP protocol **166**
Late 80s - July 1992 **167**
July 1992 - July 1994 **176**
July 1994 - today **185**
Discussion **188**
Conclusion **192**

CHAPTER 11 *Changing networks: Strategies and tools* ***193***

Introduction **193**
NORDUnet **195**
On the notion of a gateway **207**
Illustrating the roles of gateways **211**
The political incorrectness of gateways **214**

CHAPTER 1

Introduction

The emergence of information infrastructures

During the last 3 - 4 years a truly amazing transformation has taken place. From being the toy of square-eyed nerds and confined to esoteric enclaves, the interest around Internet and communicative use of IT has exploded. That advertisements would contain Web home pages, that newspapers would include a special section on Internet in addition to sports and finance and that it would be possible to tell your mother that you are “on the Net” or “surfing” is nothing short of astonishing.

In parallel with the increase in popular and media attention directed towards the Internet, the establishment of information infrastructures has been heavily promoted by political actors. The term “information infrastructure” (II) has been increasingly used to refer to integrated solutions based on the now ongoing fusion of information and communication technologies. The term became popular after the US plan for National Information Infrastructures (NII) was launched. Following that the term has been widely used to describe national and global communication networks like the Internet and more specialized solutions for communications within specific business sectors. The European Union has followed up, or rather copied, the ideas in their Bangemann report (Bangemann et al. 1994). In this way, the shape and use of information infrastructures has been transformed into an issue of industrial policy (Mansell, Mulgan XX).

The launching of ambitious political action plans like those mentioned above is to a large extent an effect of the success of the Internet - the great potential of such technologies that has been disclosed to us through the use of Internet. But the plans are over and above responses to a much broader trend concerning technological development and diffusion which the Internet also is a part of: the development of information and communication technologies. As a result of this trend, one (in particular technology policy and strategy makers) rather prefer to talk about ICT than IT or communication technologies separately. This integrated view on ICT is a result of a long term trend integrating telecommunication and information technologies - leading to a *convergence* of these technologies.

This convergence may be described as two parallel, and in principle, independent processes: information technologies “creeping into” telecommunication technologies and telecommunication technologies “creeping into” information systems. IT has crept into telecommunication first of all through the “digitalization” of telecommunication, i.e. traditional telecommunication components like switches and telephones are changed to digital technologies rather than analog technologies. The latter process has evolved as information systems have been enhanced through use of telecommunication technologies. Remote users have been given access to systems independent of physical location and independent systems have been integrated through information exchange based on services like electronic data interchange (EDI). In addition some integration processes which might be located somewhere in the middle between these two have unfolded. This include for instance the development of technologies enabling the public telephone infrastructure to be used as an interface to information systems. One example is solutions allowing bank customers to check their accounts directly using the telephone.

In total, the convergence of ICT has opened up for a vast array of new uses of technologies. The “informatization” of telecommunication has opened up for lots of new enhanced telecommunication services, and similarly, the “telecommunicatization” of information systems has opened up for an equally large range of new information systems supporting information sharing and integrating processes at a global level. The range of new solutions that seems useful and that may be developed and installed may be equally large as the number of traditional information systems developed.

Some examples of solutions of this kinds which may be available to us not too far into the future are:

- All goods might be bought through electronic commerce services. This implies that the order, inventory, invoicing and accounting systems of all companies in the world and all systems controlling customers' bank accounts in the world will be integrated.
- Global companies are integrating all their information systems globally at the same time as production control and logistics systems are integrated with all their supplies and customers,
- Individuals can access all the information services from their mobile phone: the information systems they are authorized users of, and the Internet of course - which implies that they can get access to all newspapers as well as TV channels.

The infrastructures mentioned so far can be seen as mostly been shared by the "global community." However, information infrastructures have also been developed along a different, a third, path. Inside individual corporations, the number of information systems has continuously been growing. At the same time existing systems have become increasingly integrated with each other. Most companies do not have just a collection of independent systems. The integration of and interdependence between the systems implies that they should rather be seen as an infrastructure - independent of their geographical distribution and the use of telecommunication. But this shift is of course strengthened by the fact that the companies are integrating their growing number of systems at the same time as they (drawing on the technological development sketched above) are becoming more geographically distributed (global) and integrate their systems more closely with others'.

There is also a fourth path which concerns the increasing degree ICT is penetrating and becomes embedded into our everyday lives (Silverstone and Hirsch 1992; Sørensen and Lie 1996). As individuals, we are using an increasing number of tools, systems, and services in an increasing part of our lives. ICT is not just represented in a limited set of tools we are using for specific purposes. We are rather using ICT, in one form or another whatever we are doing and wherever we are. In this way ICT is becoming a basic infrastructure for "all" our dealings in our world. We are "domesticating" the technology.

All these four trends are taking place in parallel. Just as information and communication technolgies are converging, information systems and telecommunication services are increasingly becoming linked to each other. The number of links is increasing.

Infrastructures are different

The broad trends outlined above, we argue, requires a re-conceptualization of the I(C)T solutions we are developing. If one is concerned with the development of the "information systems of the future", the phenomena we are dealing with should be seen as information *infrastructures* - not systems. This reconceptualization is required because the nature of the new ICT solutions are qualitatively different from what is captured by the concepts of information systems underlying the development of IT solutions so far. It is also significant different from traditional telecommunication systems - and even the combination of these two. This reconceptualization is what we are striving for with this book. We will here briefly motivate for the necessity of such a shift.

Understanding information infrastructures requires a holistic perspective - an infrastructure is more than the individual components. Successful development and deployment of information infrastructures requires more than a combination of traditional approaches and strategies for development of telecommunications solutions and information systems. Although these approaches complement each other, they also contain important contradictions and accordingly brand new approaches are required

The infrastructures we are addressing can still to some extent be seen as information systems as they contain everything you find in an IS. But an infrastructure is something more than an IS. In line with the brief sketch of the emergence of ICT, we can say that an infrastructure also through the "telecommunicatization" process, information infrastructures are traditional information systems plus telecommunication technology. This is of course correct, but still a bit too simple. A global infrastructure where any goods might be purchased is more than just an IS plus telecommunication. And also seen from the perspective of "informatization" of telecommunication - information is certainly more that telecom - or telecom added information systems.

Traditional approaches to information systems development are implicitly based on assumptions where the information systems are closed, stand-alone systems used within closed organizational limits. They are assumed developed within a hierarchical structure - a project (managed by a project leader and a steering group) -

which is a part of a larger hierarchical structure - the user organization (or the vendor organization in case of a commercial product). Telecommunication systems, on the other hand, are global. The most important design work - or decisions at least - are taken care of by standardization bodies (CCITT and ITU). When developing infrastructures, the focus on closed, stand-alone systems has to be replaced by one focusing on the infrastructures as open and global as is the case for development of telecommunication technologies. However, there are other parts of the telecommunication approach which is more problematic.

Characteristic for traditional telecommunication solutions have been their stability. This is in particular true for their functionality and user interface. To put it simply, the basic functionality has been stable for more than hundred years. A telephone service has one function: the user can dial a number, talk to the person at the other end, and hand up when finished. As telecommunication got “informationalized,” however, this started to change and new functions (for instance, you can transfer your phone “virtually” to another number, you may use it as an alarm clock, etc.) have been added. But the stability of the functionality is a basic precondition for how the approaches and strategies followed for developing telecommunication technologies.

What has been in focus has been the improvement of the technologies invisible to the users. At the national level, the telecommunication infrastructures have been built and operated by national monopolies since about the turn of the century (Schneider and Mayntz). Monopolies have dominated this technology as its size and interconnectedness makes this “natural,” and most investments have been made based on long time horizons, usually 30 years (ibid.). All technologies are designed and specified as global (universal) standards. Such standards have been seen as absolutely required to enable smooth operation and use of the infrastructure. However, the development of such standards takes time - usually about ten years for each standard. And under the existing conditions - the simple user interface - one (or may be three: dial, talk, hang up) operation not being modified for 100 years, the long term planning of the infrastructure, and the limited number of actors - one for each country - this approach has worked pretty well.

Information systems development, however, has very different - or rather opposite - characteristics. While the telephone functionality and user interface has been extremely stable and simple, information systems are characterized by very rich and highly dynamic functionality. Information systems are closely tied to the working processes they support. These processes are inscribed into the systems making them unique and local - not universal. The environments of information systems are highly dynamic. And the information systems are in themselves a driving force in changing the very work practices and other environmental elements they are designed to support.

In the case of telecommunication technologies the stability of the functionality and user interface has made users irrelevant in the design process. The user interface has been given, and one could concentrate only on technological issues. For information systems the situation is the opposite. As the links between the technology and the users are so rich and dynamic, dealing with the interaction between technological and human, social, and organizational issues has been of utmost importance. (This fact is not challenged by the strong technological bias of most engineers involved in IS development, struggling to turn IS development into an ordinary technical engineering discipline.)

Future information infrastructures will be just as dynamic as our information systems have been so far. They will be very heterogeneous (re. the combined vertical and horizontal integration of information systems).

So, can we find an approach to infrastructure development which account for the dynamics and non-technical elements of information systems at the same time as the standards enabling the required integration - an approach which works for more heterogeneous and dynamic technologies that telephone services up to now? This is an all to big question to be answered in one book. However, that is the question we want to inquire into. Our main focus is the tension between standardization and flexibility, which we believe is an important element in the bigger question. And in discussion this tension we will in particular be concerned about the interplay and interdependencies between technological issues on the one hand and social, organizational, and human on the other.

New trends, new research

“Everybody” observing the trends briefly described above have concluded that brand new technologies are on their way. However, large groups of IS and IT developers seem to conclude that the new technology is still software and information systems, so that good old software engineering and information systems development approaches still do.[1] Among researchers, however, the announcement of

[1] These assumptions are very visible in the methodological framework developed for engineering European wide telematics applications (..).

the Clinton/Gore and Bangemann reports triggered new activities. Several researchers soon concluded that the solutions envisions in the reports were new, and as we did not have any experience in developing such solutions, we did not have the required knowledge either (several papers in (Brandscomb and Kahin). One issue identified as the most probably key issue was that of standardization. It was considered obvious that the new infrastructures require lots of standards. Equally obvious - existing standardization practices and approaches are not at all sufficient to deliver the standards needed. An early conclusion is that the Internet experience is the most valuable source for new knowledge about this issues (Brandscomb and Kahin 1996).

Science and Technology Studies has been a rapidly growing field in the last 10-20 years. Neither standards nor infrastructures have been much in focus. However, there are a number of studies within this field which also give us important knowledge about standards useful when developing information infrastructures and which we will draw upon in this book (Geof, Marc, Leigh, ..).

IT infrastructure has also become a popular topic within information systems research focusing on the situations within individual business organizations (Broadbent, Weill XX).

Our focus, approach and contribution

II is a vast field. It covers all kinds of technologies, all kinds of use and use areas. It involves lots of political, social, organization, human aspects and issues - from the development of industrial at national, regional (EU), or even the global level within the G7 forum to the micro politics in the everyday activities between people involved in the design and use of the technology. An all these issues interact, they are interdependent and intertwined. Lots of research into all issues and all combinations are important. In this context this book will just cover a minor aspect of infrastructures. But an important one - we believe. Our focus on the standardization, and in particular the tension between standardization and flexibility. In inquiring into this tension we will in particular be concerned about the interplay between technical and non-technical (social, organizational, human, etc. (Latour,xx) issues.

We focus on *information* infrastructures. By this term we mean IT based infrastructures at the application level, not lower levels IT based telecommunication networks like for instance ATM networks or wireless networks. We see information infrastructures as “next generation” information systems. An information infrastructure can be described as an information system except that it is shared by a large user community across large geographical areas such that it might more appropriately be seen as an infrastructure than as a system.

Our motivation behind this book is to contribute to a firmer understanding of the challenges involved in establishing a working information infrastructure. Without such a basis, political action seem futile. The potential and probable impacts of information infrastructures in our everyday lives and at work, in terms of regional development and (lack of) distribution urge us to develop our ability to make informed judgements. Closing your eyes or resorting to political slogans cannot make a sound strategy.

As the information infrastructures of the future have yet to materialise — they are currently in the making — we are necessarily aimed at a moving target. Our analysis in this book is dominated by continuity: we expect and argue that the future information infrastructures will be an extension, combination, substitution and superimposition of the bits and pieces that already exist (Smarr and Catlett 1992). Hence, the experiences acquired so far is relevant. The aim of this book, then, is to paint the *emerging* picture of information infrastructures based on a critical interpretation of the fragments of existing experience.

Our intention in writing the book is addressing what we see to be (some of) the core characteristics of information infrastructures, i.e. those aspects making information infrastructures different from information systems and at the same time being critical in their development and use. These are aspects which our (research) and experience form the information systems field cannot tell us how to deal with properly.

As will be elaborated at length later, establishing a working information infrastructure is a highly involved socio-technical endeavour. Developing a firmer understanding, then, amounts to developing a reasonable account of these socio-technical processes: the actors, institutions and technologies that play a role. Seemingly, we have translated design of information infrastructures from the relatively manageable task of specifying a suitable family of technical standards and protocols to the quite unmanageable task of aligning innovations, national technology policies, conditions for commercially viable experimentation, governmental intervention and so forth. A perfectly reasonable question, then, is to inquire whether this is doable at all: is not design of information infrastructures a comforting, but ultimately naive, illusion?

Kraemer and King (1996), for instance, take a long stride in this direction in their appropriation of the NII initiative in the United States. They basically present NII as a political-economic negotiation where “there is no design (...) [only] order without design” (ibid., p. 139). Against this background, there seems to be little space for influencing the shaping of an information infrastructure through design related decisions.

Our handle on this is slightly different. We seek to make visible, not side-step, the issues of the political-economic negotiations. By making the inscriptions explicit, the intention is to pave the road for subsequent scrutiny and discussion. In this sense, we pursue a slightly, hopefully not overly, naive approach where we argue for making the most out of the available options rather than fall back into apathy or disillusion.

The fact that information infrastructures are established through complex and vast processes, implies that the notion of “designing” them needs to be critically reassessed. The connotation and assumptions about design is too much biased towards being in control of the situation. In relation to information infrastructures, we argue that it is more reasonable to think of design in terms of strategies of intervention and cultivation.

Brief outline of the book

Chapter 2 presents two examples of information infrastructures: The Internet and the information infrastructure for the Norwegian health care sector. These two cases illustrate two different types of infrastructures. They also illustrate two different development approaches. Internet is based on an evolutionary, prototype oriented approach. The Norwegian health care infrastructure is tried developed based on a specification driven approach where the focus has been on the specification of European standards for health care information exchange. These two development efforts are also different in the sense that the Internet has been growing rapidly throughout its life from the very beginning, an indisputable success. The development of the Norwegian health care infrastructure has followed a different pattern. It started with a tremendous success for one actor building the first limited network for transmission of lab reports to GPs. That solution was soon after copied by several others. After that, however, the infrastructure has developed very, very slowly.

Chapter 3 critically analyses the notion of information infrastructure. The concept is not strictly defined, but rather characterized by six key aspects. These are the aspects, we argue, making infrastructures qualitatively different from other information systems. We are arriving at these aspects by presenting and analysing a number of infrastructure definitions provided by others, including the one used in the official documents presenting the US Government plan for the building of the National Information Infrastructure. The 6 aspects are: enabling, shared, open, socio-technical, heterogeneous, and installed base.

Infrastructures have a supporting or *enabling* function. This is opposed to being especially design to support one way of working within a specific application field.

An infrastructure is one irreducible unit *shared* by a larger community (or collection of users and user groups). An infrastructure is irreducible in the sense that it is the same “thing” used by all its users (although it may appear differently), it cannot be split into separate parts being used by different groups independently. However, an infrastructure may of course be decomposed into separate units for analytical or design purposes. The fact that infrastructures are shared implies that their parts are linked and they are defined as shared standards. This means that standards are not only economically important but a necessary constituting element.

IIs are more than “pure” technology, they are rather *socio-technical networks.* This is true for ISs in general, as they will not work without support people and the users using it properly. For instance, flight booking systems do not work for one particular user unless all booked seats are registered in the systems. But this fact is largely ignored in the thinking about the design of information systems as well as infrastructures.

Infrastructures are *open.* They are open in the sense that there are no limits for the number of users, stakeholders, vendors involved, nodes in the network and other technological components, application areas, network operators, etc. This defining characteristic does not necessarily imply the extreme position that absolutely everything is included in every 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 have no connections.

Infrastructures are *heterogeneous.* They are so different in different ways. For instance, they are connected into ecologies of infrastructures as illustrated above, they are layered upon each other as in the OSI model, they are heterogeneous as they include elements of different qualities like humans and computers, etc. They are also heterogeneous in the sense that the seemingly same function might be implemented in several different ways.

Building large infrastructures takes time. All elements are connected. As time passes, new requirements appear which the infrastructure has to adapt to. The whole infrastructure cannot be change instantly - the new has to be connected to the old. In this way the old - the *installed base* - heavily influence how the new can be designed. Infrastructures are not designed from scratch, they rather evolve as the “cultivation” of an shared, open, socio-technical, heterogeneous installed base.

The remainder of the book views information infrastructures through this lens. We subsequently look deeper into the above mentioned aspects of infrastructures, and then gradually move towards design related strategies.

Based on the assumptions that standards are playing crucial roles in relation to infrastructures, chapter 4 spells out the different kinds of standards defining the Internet and the standards defined through the standardization effort organized by CEN (CEN TC/251), the organization given the authority to set European standards for health care information exchange and Internet. Different types of standards are identified and the process through which standards are worked out are described. The organisation of the standardisation process vary significantly.

The openness of infrastructures is addressed in chapter 5 and might be illustrated by an example form health care: A hospital is exchanging information with other medical institutions, even in other countries. It is exchanging information with social insurance offices and other public sector institutions, it is ordering goods from a wide range of companies, etc. These companies are exchanging information with other companies and institutions. Hospital doctors might be involved in international research programmes. Accordingly, a hospital is sharing information with virtually any other sector in society. Drawing a strict line between, for instance, a medical or health care infrastructure and an electronic commerce infrastructure is impossible. However wide an infrastructure’s user groups or application areas are defined, there will always be something outside which the infrastructure should be connected to.

Openness implies heterogeneity. An infrastructure grows by adding new layers or sub-infrastructures. Over time, what is considered to be separate or part of the same will change. Infrastructures initially developed separately will be linked together. These evolving processes make infrastructures heterogeneous in the sense that they are composed of different kinds of components.

Open worlds, like those of standards and infrastructures are dynamic and changing. To adapt to such change, infrastructures and standards must be flexible. They also need to be flexible to allow some kind of experimentation and improvement as users get experience.

Standards are neither easily made nor changed when widely implemented. Standardization means stability. The openness of infrastructures implies that the range and scope of standards must change over time, and so will their relationships to other standards. This chapter inquires into the implications of considering infrastructures open, the need for flexibility and the more intricate relationships between flexibility and change on one hand and standardization and stability on the other.

Chapter 6 outlines a theoretical framework we argue is relevant for appropriating the socio-technical aspect of information infrastructures. The framework is actor-network theory (ANT) and is borrowed from the field of science and technology studies (STS) (REF XX Latour, Bijker, Law). This chapter motivates for the relevance of ANT by comparing and contrasting it with alternative theoretical frameworks, in particular structuration theory (Orlikowski, Walsham XX). Also other scholars of information systems research have used ANT (Walsham ifp8.2, Leigh, Bloomfield et al. XX).

One key concept in actor network theory is that of “inscriptions.” This is explored at length in chapter 7. The concept explain how designers assumptions about the future use of a technology, described as programs of action, is inscribed into its design. Whether the technology in fact will impose its inscribed program of action depends on to what extent the actual program of action also is inscribed into other elements like for instance documentation, training programs, support functions, etc., i.e. into a larger network of social and technological elements (humans and non-humans).

This chapter analyses in detail what kinds of programs of action are described into two specific standardized EDIFACT message definitions for health care. We are looking at which elements, ranging from the atomic units of the message definitions to the overall organization of the standardization work, various programs of action are inscribed into as well as how these elements are aligned to each other.

So far we have inquired into the core aspects (as we see it) of information infrastructures. Starting in chapter 8, we turn more towards design related issues. In chapter 8 we look into the basic assumptions underlying most infrastructure development work, in particular standardization activities.

Most information infrastructure standardization work is based on a set of beliefs and assumptions about what a good standard is. These beliefs are strong - but are indeed beliefs as they are not based on any empirical evidence concerning their soundness. They have strong implications for what kinds of standards that are defined, their characteristics as well as choice of strategies for developing them. Beliefs of this kind are often in other contexts called ideologies. Hence, the focus of this chapter is on the dominant standardization ideology: its content, history, how it is tired applied, what really happens and its shortcomings. We will argue that it has serious short-comings. In fact, the dominant standardization approach does not work for the development of future information infrastructures. New approaches based on different ideologies must be followed to succeed in the implementation of the envisioned networks. The chapters 9 through 11 are devoted to spelling out viable, alternative design strategies to this dominating, main-stream one influenced by universalism.

The focus on infrastructure as an evolving “installed base” in chapter 9 implies that infrastructures are considered as always already existing, they are NEVER developed from scratch. When “designing” a “new” infrastructure, it will always be integrated into and thereby extending others, or it will replace one part of another infrastructure.

Within the field institutional economy some scholars have studied standards as a part of a more general phenomena labelled “self-reinforcing mechanisms” and “network externalities” (REFS XX). Self-reinforcing mechanisms appear when the value of a particular product or technology for individual adopters increases as the number of adopters increase. The term “network externalities” is used to denote the fact that such a phenomenon appears when the value of a product or technology depends also on aspects being external to the product or technology itself. Chapter 9 briefly reviews these conceptions of standards and the phenomenon we call *the irreversibility of the installed base* - the cause of this phenomenon as well as its effects. Furthermore, we look at how the irreversibility problem appears in relation to information infrastructures, its negative implications and the need for flexibility and change. This leads to a re-conceptualisation of the very notion of design of information infrastructure to something closer to a *cultivation* strategy.

The practical implications of the argument of chapter 9, that the installed base has a strong and partly neglected influence, is illustrated in chapter 10 by one specific case presented in detail. It describes the revision of the IP protocol in Internet. IP forms the core of Internet and has accordingly acquired a considerable degree of irreversibility as a result of its wide-spread use. Changes to IP have to be evolutionary and small-step, i.e. in the form of a *transition strategy.* Transition strategies illustrate the cultivation approach to design of information infrastructures.

By way of conclusion, chapter 11 explores alternative cultivation based approaches to infrastructure development. These alternatives allow more radical changes than the basically conservative character of cultivation. They are based on the use of (generalised) *gateways* which link previously unrelated networks together. In conjunction with cultivation, these gateway based approach make up the intervention strategies for information infrastructures.

Earlier papers

This book is to a large extent based on work we have reported on in earlier publications. Some material is new. And the format of a book allows us in a completely new way to combine, structure and elaborate the various threads and arguments spread out in earlier publications. These publications are listed below.

Eric Monteiro and Ole Hanseth. Social shaping of information infrastructure: on being specific about the technology. In Wanda Orlikosiki, Geoff Walsham, Matthew R. Jones, and Janice I. DeGross, editors,Information technology and changes in organisational work, pages 325 -- 343. Chapman & Hall, 1995.

Hanseth, O., Monteiro, E. and Hatling, M. 1996. Developing information infrastructure standards: the tension between standardisation and flexibility. *Science, Technology & Human Values*, 21(4):407 – 426.

Hanseth, O. 1996 Information technolgogy as infrastructure, PhD thesis, Gøteborg Univ.

O. Hanseth and E. Monteiro. Inscribing behaviour in information infrastructure standards. *Accounting, Management & Information Technologies*, 7(4):183 – 211, 1997.

O. Hanseth and E. Monteiro. Changing irreversible networks. In: Proc. ECIS ‘98, 1998.

Monteiro, E. 1998. Scaling information infrastructure: the case of the next generation IP in Internet. *The Information Society*, 14(3):229 - 245.

CHAPTER 2

Cases: The Internet and Norwegian Health Care

Throughout this book, a number of examples will be used to discuss and illustrate various issues. These examples will primary be selected from two cases - the building of two different information infrastructures: the Internet and an infrastructure for exchange of form like information in the Norwegian health care sector. The building of these infrastructures will be presented in this chapter. We also discuss methodological issues regarding how reasonable it is to draw general conclusions about information infrastructures from these cases. Our approach is pragmatic. We present an *emerging picture* of information infrastructure standardisation and development based on the empirical material at hand. This picture will be adjusted as more experience with information infrastructures is gained. The two cases exhibit, we believe, a number of salient features of information infrastructure building.

In order to make our use of the two cases as clear as possible, we identify both the important lessons to be learned as well as pointing out the more accidental, less reproducible aspects.

Internet

> "The Internet has revolutionized the computer and communications world like nothing before" (Leiner et al., 1997, p. 102).

> "The Internet today is a widespread information infrastructure, the initial prototype of what is often called the National /or Global or Galactic) Information Infrastructure" (ibid., p. 102).

As indicated by these quotes, the Internet is widely held to be the primary successful example to learn from when trying to realized the envisioned information infrastructures (Kahin and Branscomb 1995, Digital lib. 1995). We share this view, and will accordingly present the development of the Internet from its very beginning up to today. We will in this section give a brief overview of this development, pointing to what we believe to be the important steps and events that indicate what could be included in strategies for building future information infrastructures. This presentation draws heavily on (Leiner et al., 1997).

We also include a few cautionary remarks about the danger of idolizing Internet. It is a lot easier to acknowledge its historical success than to feel confident about its future.

The beginning

The first notes related to the work leading to Internet is a few papers on packet switching as a basis for computer networking written in 1962. The first long distance connections between computers based on this principles were set up in 1965. In 1969 the first nodes of ARPANET were linked together, and in 1971-72 the NCP[1] protocol was implemented on this network, finally offering network users the possibilities of developing network applications. In 1972 e-mail was introduced, motivated by the ARPANET's developers' need for easy coordination. From there, e-mail took off as the most popular network application.

The TCP/IP core

The original ARPANET grew into the Internet. Internet was based on the idea that there would be multiple independent networks of rather arbitrary design, beginning with the ARPANET as the pioneering packet switching network, but soon to include packet satellite networks, ground-based packet radio networks and other networks. The Internet as we now know it embodies a key underlying technical idea, namely that of open architecture networking. In this approach, the choice of

1. NCP, Network Control Protocol

any individual network technology was not dictated by a particular network architecture but rather could be selected freely by a provider and made to interwork with the other networks through a meta-level “Internetworking Architecture”.

The idea of open-architecture networking was guided by four critical ground rules:

- Each distinct network had to stand on its own, and no internal changes could be required of any such network before being connected to the Internet.
- Communication would be in a best-effort basis. If a packet didn’t make it to the final destination, it would quickly be retransmitted from the source.
- Black boxes (later called gateways and routers) would be used to connect the networks. No information would be retained by the gateways about individual flows of packets passing through them, keeping them simple and avoiding complicated adaptation and recovery from various failure modes.
- There would be no global control at the operations level.

The original Cerf and Kahn (1974) paper on the Internet described one protocol, called TCP, which provided all the transport and forwarding services in the Internet. Kahn had intended that the TCP protocol support a range of different transport services, from the totally reliable sequenced delivery of data (virtual circuit model) to a datagram service in which the application made direct use of the underlying network service, which might imply occasional lost, corrupted or reordered packets.

Although Ethernet was under development at Xerox PARC at that time, the proliferation of LANs were not envisioned at the time, much less PCs and workstations. The original model was national level networks like ARPANET of which only a relatively small number were expected to exist. Thus a 32 bit IP address was used of which the first 8 bits signified the network and the remaining 24 bits designated the host on that network. This assumption, that 256 networks would be sufficient for the foreseeable future, was clearly in need of reconsideration when LANs began to appear in the late 1970s.

However, the initial effort to implement TCP resulted in a version that only allowed for virtual circuits. This model worked fine for file transfer and remote login applications, but some of the early work on advanced network applications, in particular packet voice in the 1970s, made clear that in some cases packet losses should not be corrected by TCP, but should be left to the application to deal with. This led to a reorganization of the original TCP into two protocols, the simple IP which provided only for addressing and forwarding of individual packets, and the separate TCP, which was concerned with service features such as flow control and recovery from lost packets. For those applications that did not want the services of TCP, an alternative called the User Datagram Protocol (UDP) was added in order to provide direct access to the basic service of IP.

ARPANET replaced NCP with TCP/IP in 1983. This version of TCP/IP is officially given version number four.

New applications - new protocols

A major initial motivation for both the ARPANET and the Internet was resource sharing, like for instance allowing users on the packet radio networks to access the time sharing systems attached to the ARPANET. Connecting the two networks was far more economical that duplicating these very expensive computers. However, while file transfer (the ftp protocol) and remote login (Telnet) were very important applications, electronic mail has probably had the most significant impact of the innovations from that era. E-mail provided a new model of how people could communicate with each other, and changed the nature of collaboration, first in the building of the Internet itself (as is discussed below) and later for much of society.

In addition to e-mail, file transfer, and remote login, other applications were proposed in the early days of the Internet, including packet-based voice communication (the precursor of Internet telephony), various models of file and disk sharing, and early “worm” programs illustrating the concept of agents (and viruses). The Internet was not designed for just one application but as a general infrastructure on which new applications could be conceived, exemplified later by the emergence of the Web. The general-purpose nature of the service provided by TCP and IP makes this possible.

As TCP moved to new platforms, new challenges were met. For instance, the early implementations were done for large time-sharing systems. When desktop computers first appeared, it was thought by some that TCP was too big and complex to run on a personal computer. However, well working implementations were developed, showing that such small computers could be connected to Internet as well.

As the number of computers connected increased, new addressing challenges appeared. For example, the Domain Name System was developed to provide a scalable mechanism for resolving hierarchical host names (like ifi.uio.no) into Internet addresses. The requirements for scalable routing approaches led to a hierarchical model of routing, with an Interior Gateway Protocol (IGP) used inside each region of the Internet and an Exterior Gateway Protocol (EGP) used to tie the regions together.

Diffusion

For quite some time, the Internet (or at that time ARPANET) was primarily used by its developers and within the computer networking research community. As the next major step, the technology was adopted by groups of computer science researchers. An important use area was the development of basic support services for distributed computing in environments based on work stations connected to LANs like distributed (or networking) file systems. A crucial event in this respect was the incorporation of TCP/IP into the Unix operating system. An additional important element was the fact that the code was freely distributed. And lastly, the on-line availability of the protocols' documentation.

In 1985 the Internet was established as a technology supporting a broad community of researchers and developers.

The evolution of the organization of the Internet

> "The Internet is as much a collection of communities as it is a collection of technologies" (ibid. p. 106).

It started with the ARPANET researchers working as a tight-knit community, the ARPANET Working Group (Kahn 1994). In the late 1970s, the growth of the Internet was accompanied by the growth of the interested research community and accordingly also an increased need for more powerful coordination mechanisms. The role of the government was initially to finance the project, to pay for the leased lines, gateways and development contracts. In 1979, one opened up for participation from a wider segment of the research community by setting up Internet Configuration Control Board (ICCB) to overlook the evolution of Internet. ICCB was chaired by ARPA (Kahn 1994, p. 16). In 1980 the US Department of Defence adopted TCP/IP as one of several standards. In 1983 ICCB was substituted by the IAB, Internet Activities Board. The IAB delegated problems to task forces. There were initially 10 such task forces. The chair of the IAB was selected from the research community supported by ARPA. In 1983 TCP/IP was chosen as the standard for Internet, and ARPA delegated the responsibility for certain aspects of the standardisation process to the IAB.

During the period 1983 -- 1989 there is a steady growth in the number of task forces under the IAB. This gives rise to a reorganisation in 1989 where the IAB is split into two parts: (i) the IETF (Internet Engineering Task Force) which is to consider “near-future” problems and (ii) IRTF (Internet Engineering Research Task Force) for more long term issues.

The Internet is significantly stimulated by the High Performance Computing initiate during the mid-80s where a number of supercomputing centres were connected by high-speed links. In the early 90s the IAB is forced to charge nominal fees from its members to pay for the growing administration of the standardisation process.

In 1992 a new reorganisation took place. Internet Society was established as a professional society. IAB got constituted as part of the Internet Society. While keeping the abbreviation, the name was change from Internet Activities Board to Internet Architecture Board. IAB delegated the responsibility for the Internet Standards to the top level organization within the IETF which is called IESG (Internet Engineering Steering Group). The IETF as such remained outside of the Internet Society to function as a “mixing bowl'” for experimenting with new standards.

The Web’s recent development and widespread development brings in a new community. Therefore, in 1995, a new coordination organization was formed - the World Wide Web Consortium (W3C). Today, the W3C is responsible for the development (evolution) of the various protocols and standards associated with the Web. W3C is formally outside the IETF but is closely related. This is partly due to formal arrangements but more important is the similarity between the standardisation processes of the two (see chapter 11), and the fact that many of those active in W3C have been involved in other Internet activities for a long time, being members of the larger Internet community and sharing the Internet culture (Hannemyr 1998, ++).

Future

Internet has constantly been changing, seemingly at am increasing speed. The outphasing of IP version 4, adopted by the whole ARPANET in 1983, is now at its beginning. The development of the new one, IP version 6, started in 1991 and was made a draft standard in 1996 (see further explanation in chapter 4). The transition of the Internet to the new one is at its very beginning. The details of this evolution is an important and instructive case of changing a large infrastructure and will be described in detail later in chapter 10.

The development of this new version turned out to be far more complicated than anticipated - both technologically and organizationally. Technologically due to the complexity of the Internet, organizationally due to the number of users and user groups involved. The Internet is supposed to be the underlying basis of a wide range of new services, from new interactive media to electronic commerce. To play this future role the Internet has to change significantly. New services such as real-time transport, supporting, for instance, audio and video streams have to be provided: It also has to be properly adapted to new lower level services such as broadband networks like ATM and Frame Relay, portable computers (lap tops, PDAs, cellular phones) and wireless networks enabling a new paradigm of nomadic computing, etc. The required changes confront us with challenging technological as well as organizational issues.

> "The most pressing question for the future of the Internet is not how the technology will change, but how the process of change and evolution itself will be managed" (Leiner at al., p. 108).

Highlights

We will here point to what we consider the key lessons to be learned from the Internet experience (so far) and which will be focused throughout this book.

Fist of all, the Internet's has constantly changed. It has changed in many ways, including:

1. The Internet's character has changed.

   It started as a research project, developing new basic network technologies. As it developed, it also became a network providing services for its developers, then a network providing services to researcher communities, and lastly a network supporting all of us. As the network changed, its organization had to adapt - and it has done so.
2. It has constantly been growing.

   It has grown in terms of number of nodes (hosts and routers) connected, in number of users and use areas, and in terms of protocols and applications/services.
3. The Internet has had to change.

   It has had to change to accommodate to its own growth as well as its changing environment. Among the first changes are the introduction of the Domain Name System and the change from initial 32 bit address to the one used in IPv4, and now the definition of IPv6 to allow continued growth. Among the latter are changes necessary when PCs and LANs were developed and diffused and the above mentioned ongoing changes to adapt to broadband and wireless networks etc.

Basic principles for the development has been:

1. Establishing a general basis for experimental development of applications and services. This general basis was a packet switched computer communications network, itself being subject to experimentation.
2. Applications and services have been implemented to support specific local needs. Widely used services are based on applications that turned out to be generally useful. E-mail and WorldWideWeb are examples of applications originally developed for such specific local needs.
3. When an application has proved to be of general interest, its specification is approved as standard.
4. The Internet was always been an heterogeneous network.

   It has been heterogeneous as it from its very beginning was designed to integrate, run across, various basic networks like telephone, radio, satellite, etc. It has also been heterogeneous by accepting two alternative protocols, TCP and UDP, on the same level. Today is the Internet also heterogeneous as it has integrate various different network on higher levels like America On-Line, prodigy, etc. with its own protocols and services an e-mail networks based on other protocols like X.400, cc:mail, etc.[2]

Historical coincidences

Given the present, overwhelming success of Internet, there is a pronounced danger that this might tend towards idealizing the Internet experience. It is important to develop a sense of how far the Internet experience is relevant as a basis for generalized lessons, and what should be regarded as more or less historically contingent,

2. This is a somewhat controversial standpoint. The Internet community has always stressed that “reality” out there is heterogeneous, and accordingly a useful running network has to run across different basic network technologies. This has been a crucial argument in the “religious war” against OSI. However, the Internet community strongly believe in “perfect” technical solutions (Hannemyr 1997), and accordingly refuse to accept gateways between not perfectly compatible solutions (Stefferud 199x). We return to this when discussing gateways in chapter 11.

irreproducible decisions. Internet has, of course, a number of historically contingent features which distinguish it and make it difficult, if not impossible, to intentionally reproduce. Among these are:

- For a long time the developers were also the users.

  This was important for the experimental development and early use of new services. These factor can hardly be replicated when developing services for, say, health care and transport.
- The inclusion of the technology into Berkley Unix.
- The free distribution of protocol implementation ins general as well as Berkley Unix including the Internet technology in particular.
- The development of killer applications, in particular e-mail and World-WideWeb.
- The small boy's club atmosphere which prevailed during the early years, right up till today, was important for the way the work was organized and spread.

In addition, as in all cases, there has been a number of coincidences where independent events have happened at the same time, opening up possibilities and opportunities creating a success story.

International Standardization Organization (ISO)

ISO has a long history of developing standards in all kinds of areas, ranging from the length of a meter to nuts and bolts (REF). Due to the way it is intended to mimic quasi-democratic decision processes with representative voting (see chapter 4 for further details), ISO has an unique position within national technology policy. Member nations agree to make ISO standards official, national standards. This implies that ISO standards are automatically promoted by key, public actors in the areas they cover. And in some countries (like Norway), ISO standards are granted the status of laws.

When the need for open (i.e. non-proprietary) computer communication standards was gaining wide acceptance, ISO was a quite obvious choice of body being responsible for the standardization work. The development of the OSI model and its protocol suite, covering everything from coding of physical signals to applications like e-mail and secure transactions, started in XXXX. The standards are specified in terms of a so-called reference model, the Open Systems Interconnection

(OSI) model. It specifies seven layers which in sum make up what was supposed to become the basis of what we call information infrastructures.

For several years the was a “religious war” between the Internet and OSI supporters (ref.). Beyond the religious aspects, the Internet and OSI work reflected different visions about what our future world of computer communications should look like, giving different actors different roles. Accordingly, the battles over technological design alternatives, were also battles over positions in future markets (Abbate 1995). In addition, the war had a general political content as the Internet technology was developed by Americans, and accordingly giving them an competitive advantage. For the Europeans, then, it was important to make OSI different from the Internet technology, putting them in an equal position.

The OSI approach was indeed different. It negates virtually every element in the list of “highlights” above. The protocols have been tried developed by everybody coming together, agreeing on the protocols specification. No experimentation, no implementation before standardization, no evolution, no change, no heterogeneity.

Now the war is over, although there might still be some lonely soldiers left in the jungle to whom this message still has not reached.[3]

EDI and EDIFACT

One important form of computer communication is EDI)Electronic Data Interchange). This form of communication covers the exchange of information between different organizations, typically information having been exchanged as paper forms, often even formally standardized forms. Such forms include orders, invoices, consignment notes and freight bills, customs declaration documents, etc. In the business world, EDI infrastructures have been built for quite some time, and bodies taking care of the standardization have been set up. The seemingly largest and important activity is related to EDIFACT.[4] EDIFACT is a multifaceted creature. It is a format, or language, for defining data structures combined with rules for

3. OSI has been strongly supported by public agencies, and most governments have specified their GOSIPs (Government OSI Profiles), which have been mandatory in government sectors. These GOSIPs are still mandatory, and Internet protocols are only accepted to a limited extent.

how such structures should be encoded into character streams to be exchanges. Further, it includes a set of standardized data structures and a large bureaucratic organization controlling the EDIFCAT standards and standardization processes.

The primary items to be standardized in the EDIFACT world are “messages.” An EDIFACT message is typically an electronic equivalent of a paper form.

The EDIFACT standardization organization is a part of the United Nations system. This was a deliberate choice by those starting the EDIFACT standardization work, believing that United Nations would give the activities the best possible legitimation.

The EDIFACT format was defined in the early seventies, while the formal standardization activities started in XX.

Health care information infrastructures

Health care information infrastructures have been developed and used over a period of more than ten years and have taken different shapes over time. Two main forms of use are transmission of form-like information and (possibly real-time) multi-media information. Typical examples of the former include: lab orders and reports exchanged between general practitioners, hospitals and labs; admission and discharge letters between general practitioners, specialists, and hospitals; prescriptions from general practitioners to pharmacies; exchange of non-medical information like ordering of equipment and food and invoices from hospitals and general practitioners to health insurance offices for reimbursement.

Typical examples of the latter type include: telemedicine services, that is, computer based services which usually include real time multi-media conferencing systems supporting a physician requesting advise from another physician at another institution; access to data bases and Web servers containing medical information; and PACS (Picture Archive Systems for X-rays) systems. In this book, we focus on the former type, i.e. transmission of form-like information.

[4] EDIFACT is an abbreviation for Electronic Data Interchange For Administration, Commerce and Transport.

The various forms of information exchange are overlapping and interconnected. The same piece of information may be exchanged as part of different transactions, for instance, by transmission of a digital X-ray image either using a multi-media conference system or attached in an e-mail. Furthermore, any organisational unit may engage in transactions with several other units. A lab, for instance, may communicate with a number of general practitioners, other labs, and other hospital wards. This is what distinguish such systems from stand-alone applications and turn them into infrastructure.

The development of health information infrastructure standards — not to mention their implementation in products and adoption by user organisations — has been slow. Based on personal experience, it seems that the more formal the standardisation process is, the slower the adoption becomes. Industrial consortia seem so far to be most successful. As, to the best of our knowledge, there does not exist any systematic evaluation, this is difficult to "prove." But studies in other sectors than health care exist. The evaluation of the European Union's program for diffusion of EDI in trade, the TEDIS programme, lend support to the view that formal standardisation is incredible slow - design as well as diffusion (Graham et al. 1996). An evaluation of EDIFACT on behalf of the United Nations concludes similarly (UN 1996).

EDI Infrastructure in the Norwegian health care sector

Fürst

The development of electronic information exchange between health care institutions in Norway started when a private lab, Dr. Fürst's Medisinske Laboratorium in Oslo, developed a system for lab report transmission to general practitioners in 1987. The system was very simple — the development time was only 3 weeks for one person (Fiskerud 1996). The interest of Dr. Fürst's laboratory was simply to make profit by attracting new customers. It was based on the assumption that the system would help general practitioners save much time otherwise spent on manual entry of lab report data, and that the general practitioners would find the possibility of saving this time attractive. Each general practitioner receives on average approximately 20 reports a day, which take quite some time to register manually in their medical record systems.

The system proved to be a commercial success and brought them lots of general practitioners as new customers. This implied less profit for the other labs. Within a couple of years, several non-private labs (in hospitals) developed or bought systems with similar functionality in order to be competitive. Alongside the growing number of labs adopting systems for exchange of reports, an increasing number of actors saw a wider range of applications of similar technology in other areas. These actors were represented within the health sector as well as among possible vendors of such technology. For all of them it was perceived as important that the technologies should be shared among as many groups as possible in order to reduce costs and enable interconnection of a wide range of institutions.

The network Fürst established is still in use. Currently Fürst delivers the reports to about 2000 regular customers through the network.

Telenor - standardization

After an initial experiment, Telenor (the former Norwegian Telecom) launched the project “Telemedicine in Northern Norway” in 1987 which was running until 1993. Standardisation has always been considered important within the telecommunication sector. This attitude combined with Telenor’s interest in largest possible markets, made them take it for granted that the new health information infrastructure standards should be like any other telecommunication standard: “open” and developed according to the procedures of formal standardisation bodies.

As there was a general consensus about the need for standards, the fight about what these standards should look like and how they should be developed started. Based on their interests in general solutions and rooted in the telecommunication tradition of international standardisation, Telenor searched for international activities aiming at developing “open” standards. The IEEE (Institute of Electrical and Electronics Engineers) P1157 committee, usually called Medix, did exactly that. This work was the result of an initiative to develop open, international standards taken at the MEDINFO conference in 1986. Medix, which was dominated by IT professionals working in large companies like Hewlett Packard and Telenor and some standardisation specialists working for health care authorities, adopted the dominating approach at that time, namely that standards should be as open, general and universal as possible.

The idea of open, universal standards underlying the Medix effort implied using existing OSI (Open Systems Interconnection) protocols defined by the ISO (International Standardisation Organisation) as underlying basis. The Medix effort adopted a standardisation approach — perfectly in line with texts books in information systems development — that the development should be based on an information model being a “true” description of the relevant part of reality, that is, the health care sector, independent of existing as well as future technology. (More on this in later chapters, particularly 8 and 12.) Individual messages would be derived from the model more or less automatically.

While the focus was directed towards a comprehensive information model, lab reports were still the single most important area. However, for those involved in Medix the task of developing a Norwegian standardised lab report message had around 1990 been translated into the development of a proper object-oriented data model of the world-wide health care sector.

In addition to the information model, protocols and formats to be used had to be specified. In line with the general strategy, as few and general protocols and formats as possible should be included. Medix first focused on the ISO standard for exchange of multi media documents, ODA (Open Document Architecture) believing it covered all needs for document like information. However, around 1990 most agreed that EDIFACT should be included as well. The Europeans who strongest advocated EDIFACT had already established a new body, EMEDI (European Medical EDI), to promote EDIFACT in the health sector. In Norway, a driving force behind the EDIFACT movement was the “Infrastructure programme” run by a governmental agency (Statskonsult) during 1989 - 92. Promoting Open Systems Interconnection standards and EDIFACT systems based on Open Systems Interconnection were key goals for the whole public sector (Statskonsult 1992).

Several other standardization activities were considered and promoted/advocated by various actors (vendors). Andersen Consulting, for instance, promoted the HL-7[5] standard. The Ministry of Health hired a consultant making their own proposal - which immediately were killed by other actors on this arena. The dispute settled in 1990 when the Commission of the European Community delegated to CEN (Comite Europeen de Normalisation, the European branch of ISO) to take responsibility for working out European standards within the health care domain in order to facilitate the economical benefits of an European inner market. CEN established a so-called technical committee (TC 251) on the 23. of March 1990 dedicated to the development of standards within health care informatics. From this time Medix disappeared from the European scene. However, the people involved moved to CEN and CEN’s work to a large extent continued along the lines of Medix.

5. HL-7 is an abbreviation of Health Level 7, the number “7” referring to level 7 in the OSI model.

Both CEN and Medix were closely linked to the OSI and ISO ways of thinking concerning standards and standardization: Accordingly, the work has been based on the same specification driven approach - and with the seemingly same lack of results.

Isolated networks for lab report exchange

As mentioned above, a large number of labs bought systems similar to Fürst's. Some of these were based on early drafts of standardized EDIFACT messages, later update to the standardized versions. These networks were, however, not connected to each other. They remained isolated islands - each GP connected could only receive report from one lab.

Each of the networks got a significant number of GPs connected within a short period after the network technology was put in place. From then on, the growth has been very, very slow.

Pilots

In other areas, a number of pilot solutions have been implemented and tested in parallel with the standardization activities. This has been the case for prescriptions, lab orders, physicians' invoices, reports from X-ray clinics and other labs like pathology and micro-biology, etc. However, non of these systems has been adopted in regular use.

Cost containment in the public sector

All activities mentioned above has been driven from an interest in the improvement of medical services. Some overlapping activities were started under Statskonsult´s "Infrastructure programme." The overall objectives of this programme was to improve productivity, service quality, and cost containment in the public sector.

It is widely accepted that physicians get too much money for their work from the social insurance offices. Through improved control the government could save maybe more than hundred billion Norwegian kroner a year.[6] Spendings on pharmaceuticals are high, and is accordingly another important area for cost containment. In addition, the health care authorities wanted enhanced control concerning the use

6. These numbers are of course controversial and debatable.

of drugs by patients as well as prescription practices of physicians concerning habit-forming drugs.

As part of the Infrastructure programme KITH (Kompetansesenteret for IT i helsesektoren) was hired by Statskonsult to work out a preliminary specification of an EDIFACT message for prescriptions (KITH 1992). A project organization was established, also involving representatives for the pharmacies and the GPs.

The interests of the pharmacies were primarily improved logistics and eliminating unnecessary retyping of information (Statskonsult 1992). By integrating the system receiving prescriptions with the existing system for electronic ordering of drugs, the pharmacies would essentially have a just-in-time production scheme established. In addition, the pharmacies viewed it as an opportunity for improving the quality of service to their customers. A survey had documented that as much as 80% of their customers were favourable to reducing waiting time at the pharmacies as a result of electronic transmission of prescriptions (cited in Pedersen 1996).

As the standardization activities proceeded, the project drifted (Ciborra 1996, Berg 1997 a, b) away from the focus on cost containment. The improved logistics of pharmacies became the more focused benefit. The technological aspects of the standardization work contributed to this, as such an objective appeared to be more easy to obtain through standard application of EDI technology.

This drifting might have been an unhappy event, as the potential reduction in public spending could significantly help raising the required funding for developing a successful system.

Highlights

We will here, as we did for the Internet, point to some lessons to be learnt from this experience:

1. Simple solutions being designed and deployed under a strong focus an the specific benefits to be obtained have been very successful.
2. Health care is a sector with a wide range of different overlapping forms of communication, and communication between many different overlapping areas.
3. Focusing on general, universal standards makes things very complex, organizationally as well as technologically. The specification of the data format used by Fürst takes one A4 page. The specification of the CEN standardized EDIFACT message for lab reports takes 500 (!) pages (ref. CEN). Where the CEN message is used, the date being exchanges are exactly the same as in the Fürst solu-

tion. The focus on general solutions also makes the benefits rather abstract, and the solutions are hard to sell to those who has the money. Furthermore, there is a long way to go from such general standards to useful, profitable solutions.

4. A strong focus on standards makes success unlikely.

Methodological issues

As pointed above and elaborated further in chapter 3, there is a wide variety of information infrastructure standards produced within bodies ISO/CEN, EDIFACT, and Internet Society. These standards are on different levels and deals with issues like message definitions, syntax specification, protocols, file type formats, etc. Some standards are general purpose, others are sector specific ones (for instance, health care), and some are global while others are regional. Most of them are currently in-the-making. Our study does not provide evidence for drawing far-reaching conclusions regarding all types of information infrastructure standards. We believe, however, that the health care standards we have studied are representative for crucial parts of the standards of the information infrastructures envisioned in for instance the Bangemann report (1994), and that the picture of standardisation emerging from our analysis contains important features.

Studying the development of information infrastructures is not straightforward. There are at least two reasons for this which have immediate methodological implications worth reflecting upon.

First, the size of an information infrastructure makes detailed studies of all elements practically prohibitive. Internet, for instance, consists of an estimated 10 million nodes with an unknown number of users, more than 200 standards which have, and still are, extended and modified over a period of 25 years within a large, geographically dispersed organisation where also a number of vendors (Sun, IBM, Microsoft), commercial interests (MasterCard, Visa) and consortia (W3) attempt to exercise influence. This implies that the notion of an actor in connection with information infrastructure standardisation is a fairly general one in the sense that it is sometimes an individual, a group, an organisation or a governmental institution. An actor may also be a technological artifact — small and simple or a large system or network like Internet or EDIFACT.

Actor network theory has a scalable notion of an actor as Callon and Latour (1981, p. 286) explain: “[M]acro-actors are micro-actors sitting on top of many (leaky) black boxes”. In other words, actor network theory does not distinguish between a macro- and micro-actor because opening one (macro) black-box, there is always a new actor-network. It is not a question of principle but of convenience, then, which black-boxes are opened and which are not. To account for information infrastructure standardisation within reasonable space limits, it is necessary to rely on such a scalable notion of an actor. A systematic, comprehensive empirical study is prohibitive. In our study, we have opened some, but far from every, black-box. Several black-boxes have been left unopened for different reasons: some due to constraints on writing space, some due to lack of data access and some due to constraints on research resources. It might be desirable to have opened more black-boxes than we have done. We believe, however, we have opened enough to be able to present a reasonable picture of standardisation.

Our empirical evidence is partly drawn from standardisation of EDI messages within the health care sector in Norway. A method of historical reconstruction from reports, minutes and standards documents together with semi- and unstructured interview has been employed, partly based on (Pedersen 1996). One of the authors was for a period of three years engaged in the development of the standards by two of the companies involved (Telenor and Fearnley Data). Our account of the case has been presented, discussed and validated with one of the key actors (KITH, Norwegian: Kompetansesenteret for IT i Helsesektoren A/S).

Second, the fact that information infrastructures are currently being developed and established implies that there is only limited material on about the practical experience with which solutions “survive” and which “die”, i.e. which inscriptions are actually strong enough to enforce the desired pattern of use. Hence, we are caught in a dilemma. On the one hand, the pressure for grounding an empirical study suggests that we need to await the situation, let the dust settle before inquiring closer. On the other hand, we are strongly motivated by a desire to engage in the ongoing process of developing information infrastructures in order to influence them (Hanseth, Hatling and Monteiro 1996).

Methodologically, the use of Internet as a case, in particular the IPng case described in chapter 10, is a historical reconstruction based on several sources. Internet keeps a truly extensive written record of most of its activities, an ideal source for empirical studies related to the design of Internet. We have used the archives for (see chapter 4 for an explanation of the abbreviations) IETF meetings including BOFs, working group presentations at IETF meetings (ftp://ds.internic.net/ietf/ and http://www.ieft.org), RFCs (ftp://ds.internic.net/rfc/), minutes from

IPng directorate meetings (ftp://Hsdndev.harvard.edu/pub/ipng/directorate.minutes/), e-mail list for big-internet (ftp://munnari.oz.au/big-internet/list-archive/) and several working groups (ftp://Hsdndev.harvard.edu/pub/ipng/archive/), internet drafts (ftp://ds.internic.net/internet-drafts/), IESG membership (http://ietf.org/iesg.html#members), IAB minutes (http://info.internet.isi.edu:80/IAB), IAB membership (http://www.iab.org/iab/members.html) and information about IAB activities (http://www.iab.org/iab/connexions.html). The archives are vast, many thousand pages of documentation in total. The big-internet e-mail list, for instance, receives on the average about 200 e-mails every month. As a supplement, We have conducted in-depth semi-structured interviewing lasting about two hours with two persons involved in the design of Internet (Alvestrand 1996; Eidnes 1996). One of them is area director within IETF and hence a member of the IESG.

CHAPTER 3

Defining information infrastructures

Introduction

An important part of the problem with information infrastructures (IIs) is exactly how to look at them: what kind of creatures are they really, what are the similarities and differences compared to other classes of information systems, how should IIs be characterised, is it possible to draw a clear line between IIs and other information systems or is this rather a matter of degree, perspective or, indeed, inclination? And assuming that it is reasonable to talk about IIs, why do they — and should they, we want to argue — attract attention from media, politicians and scholars? These are highly relevant, but difficult, questions we address in this chapter.

The term II has been widely used only during the last couple of years. It gains its rhetorical thrust from certain so-called visions. These visions were initiated by the Gore/Clinton plans and followed up by the European Union's plan for Pan-European II. The visions for an II are argued as a means for "blazing the trail (...) to launch the information society" (Bangemann et al. 1994, 23). The Bangemann commission proposed ten applications which this effort should be organised around within the European Union: teleworking, distance learning, university and research networks, telematics services for small and medium sized enterprises, road traffic management, air traffic control, health care networks, electronic tendering, trans-European public administration network and city information highways. The proposal is in line with the projects proposed by the Group of seven (G7) in Brussels in March 1995.

The Bangemann commission (ibid.) states that by building a European IIs we can expect

- A more caring European society with a significantly higher quality of life and a wider choice of services and entertainment for their citizens and consumers.
- New ways to exercise creativity for content providers as the information society calls into being new products and services.
- New opportunities for European regions to express their cultural traditions and identities and, for those standing on the geographical periphery of the Union, a minimising of distance and remoteness.
- More efficient, transparent and responsive public services, closer to the citizen and at lower cost.
- Enterprises will get a more effective management and organisation of enterprises, improved access to training and other services, data links with customers and suppliers generating greater competitiveness.
- Europe's telecommunications operators will get the capacity to supply an ever wider range of new high value-added services.
- The equipment and software suppliers; the computer and consumer electronics industries: will get new and strongly-growing markets for their products at home and abroad.

Less speculative than citing political manifestoes, it is fairly safe to expect that future II will consist of an elaboration, extension and combination of existing computer networks with associated services (Smarr and Catlett 1992). It is likely to consist of an inter-connected collection of computer networks, but with a heterogeneity, size and complexity extending beyond what exists today. New services will be established, for instance, by developing today’s more experimentally motivated services for electronic commerce, video-on-demand and electronic publishing. These new services subsequently accumulate pressure for new development of the II to accommodate them.

There exist today a number of embryonic manifestations of the IIs. For many years, we have had application specific networks. Services provided include flight booking and bank networks supporting automatic teller machines and other economic transactions. Electronic data interchange (EDI), that is, electronic transmission of form-like business and trade information, is an illustration of an existing technology related to II (Graham et al. 1995; Webster 1995). The rapid diffusion of World-

WideWeb is the basis of a general II for information exchange as well as more specialised IIs implementing open electronic marketplaces where products may be ordered, paid for and possibly delivered (if they exist in electronic form like books, newspapers, software or stock marked information).

Basic data communication technology may be described as communication standards and the software and hardware implementing them. This technology is in many respects what comes closest to an existing, general purpose II. Two such basic communication technologies exist, Open Systems Interconnection (OSI) and Internet (Tanenbaum 1989). OSI is developed by the International Standardization Organization (ISO).

There is today no clear-cut conception of what an information infrastructure is and even less how to design one. We approach this in a somewhat indirect fashion. We trace the origin of the concept, discuss proposed definitions, compare it with other infrastructure technologies and outline the role and contents of standards.

Roots

The concept of II may be seen as a combination, or merge, of information and infrastructure technologies. IIs can be seen as a step in the development of information technologies as well as a step in the development and infrastructure technologies. IIs share a number of aspects with other kinds of information technologies while having some unique aspects making them different. The term "infrastructure" has been used in relation to information technology to denote basic support systems like operating systems, file servers, communication protocols, printers, etc. The term was introduced to separate between such underlying support services and the applications using them as the complexity of computing in organizations rose. The II examples mentioned above can also be seen as an evolution of computer networks, interorganizational systems and distributed information systems. IIs as envisioned are similar to examples of these concepts except that they are larger, more complex, more diversified, etc. It accordingly makes betters sense to talk about an information systems as having degrees of infrastructural aspects.

Various forms of infrastructure like railways, roads, telecommunication networks, electricity supply systems, water supply, etc. have been analysed under the label "large technical systems" (Mayntz and Hughes 1988, La Porte 1989, Summerton 1994). Some scholars have focused on what they find to be important characteristics of such technologies, talking about networking (David 1987, Mangematin and Callon 1995) and systemic technologies (Beckman 1994).

Our approach to the study of the characteristics of IIs is to focus on what is found to be the primary characteristics of other infrastructure technologies in general and analyse how these characteristics appear in IIs.

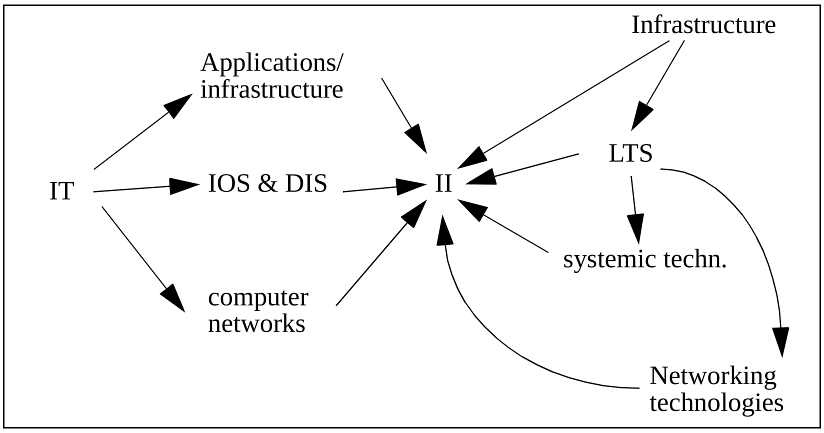

FIGURE 1. The conceptual heritage of IIs

Defining II - identifying the aspects

We will now identify what we consider to be the key aspects of (information) infrastructures, and in particular what makes them different from information systems. These aspects will be identified by presenting and discussion a number of definitions proposed by others.

Enabling, shared and open

In Webster's dictionary infrastructure is defined as

> "a substructure or underlying foundation; esp., the basic installations and facilities on which the continuance and growth of a community, state, etc. depends as roads, schools, power plants, transportation and communication systems, etc." (Guralnik 1970).

This definition is also in perfect harmony with how IIs are presented in the policy documents mentioned above. Some elements may be emphasized, leading to the identification of the first three core aspects:

> Aspect 1: Infrastructures have a supporting or *enabling* function.

This means that an infrastructure is design to support a wide range of activities, 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 improving or automating something existing. This is opposed to being especially design to support one way of working within a specific application field. This enabling feature of infrastructures plays important roles in policy documents like those mentioned above. The enabling and constraining character of II technologies will be discussed in chapter 7.

> Aspect 2: 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 one and 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 being used by different groups independently. An e-mail infrastructure is one such shared irreducible unit, while various installation of a word processor may be used completely independently of each other. However, an infrastructure may of course be decomposed 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, is all too expensive. As we see it, 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 no real infrastructure, but just a collection of independent connections. We return to this in the next chapter.

The shared and enabling aspects of infrastructures have made the concept increasingly popular the later years. Just as in the case of IIs 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 ship building sector has stayed competitive through major changes from sailing ships, through steam boats and tankers to offshore oil drilling supply boats due to the crucial role of an infrastructure of competence centres and supporting institutions like shipbuilding research centres, a ship classification company and specialized financial institutions. In the same way, Ed Steinmueller (1996) illustrate how the shared character of infrastructures is used to help understand the growing importance of knowledge as public goods and infrastructure in societies where innovation and technological development is crucial for the economy. Different II standards and standardization processes will be presented in chapter 4. However, aspects of standards will be an underlying theme through the whole book.

Aspect 3: Infrastructures are *open.*

They are open in the sense that there is no limits for number of user, stakeholders, vendors involved, nodes in the network and other technological components, application areas or network operators. This defining characteristic does not necessarily imply the extreme position that absolutely everything is included in every II. 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 have no important or relevant connections.

This might be illustrated by an example form health care: A hospital is exchanging information with other medical institutions, even in other countries. It is exchanging information with social insurance offices and other public sector institutions and it is ordering gods from a wide range of companies. These companies are exchanging information with other companies and institutions etc. Hospital doctors might be involved in international research programmes. Accordingly, a hospital is sharing information with virtually any other sector in society. And the information exchanged among different partners is overlapping. Drawing a strict line between, for instance, a health care and an electronic commerce infrastructure is impossible. However wide an infrastructure's user groups, application ares, designers and manufacturers, network operators or service providers are defined, there will always be something outside which the infrastructure should be connected to.

Unlimited numbers of users, developers, stakeholders, components and use areas implies

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

In sum - all this implies heterogeneity

Unix systems and networks based on OSI protocols are closed according to this definition although they are declared open by their name. The openness of IIs will be discussed further in chapter 5.

Heterogeneity

In the Clinton/Gore report, the envisioned NII 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:

- A wide range 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, etc.
- *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.

It is said that every component of the information infrastructure must be developed and integrated if America is to capture the promise of the Information Age.

The Bangemann report does not contain any definition. However, the report is by and large a European response to - or rather a copy of - the US NII report. It seems to share all its basic assumptions, and accordingly we can say that it implicitly also use the same II definition.

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 along many different dimensions.

> Aspect 4: IIs are more than “pure” technology, they are rather *socio-technical networks.*

Infrastructures are heterogeneous concerning the qualities of their constituencies. They encompass technological components, humans, organizations, and institutions. This fact is most clearly expressed in the last bullet paragraph above. This is true for information technologies in general, as they will not work without support people. An information system does not work either if not the users are using it properly. For instance, flight booking systems do not work unless all booked seats are registered in the systems.

The socio-technical aspects of IIs will be explored and discussed in chapter 6 and chapter 7.

> Aspect 5: Infrastructures are connected and interrelated, constituting *ecologies of networks.*

They are so different in different ways. For instance, different technologies are brought together as illustrated in the NII definitions. Here we will concentrate on different ways (sub-)infrastructures are connected into ecologies of infrastructures. Infrastructures are

- layered;
- linking logical related networks; and
- 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.

Infrastructures are also heterogeneous in the sense that the same logical function might be implemented in several different ways. We will discuss heterogeneity being caused by two forms of infrastructure development:

1. 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 phenomena is the transition of Internet from IPv4 to IPv6 (Hanseth, Monteiro and Hatling, 1996, Monteiro 1998, chapter 10).
2. Larger infrastructures will often be developed by interconnecting two existing different ones, as typically has happened when networks like America Online and Prodigy have been connected to Internet through gateways.

The third form of heterogeneity we are addressing is what happens when 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 change for whatever reason, this will often require the others to be changed as well. Examples of this phenomena are various formats for representing text, video, sound, image and graphical representations are brought together and put into MIME to enable transfer of multimedia information on the Web/Internet.

Different networks — some compatible and closely aligned, others incompatible and poorly aligned — are superimposed, one on top of the other, to produce an ecology of networks. The collection of the different, superimposed communication infrastructures in a city provides a vivid illustration of the situation: the different telephone operators, the mobile phones, data communication, the cable-TV networks have to a lesser or larger extent evolved independently but are getting more and more entagled (REF BYPLAN BOKA). We return to and elaborate on this aspect of an II in chapter 11.

While heterogeneity is indeed present in the NII definition of II, this seems not to be the case in mere engineering inspired definitions. We see the following definition proposed by McGarty (1992, p. 235-236) to be representative for communities designing computer communication technologies. He says that an infrastructure resource is

- *Shareable*. The resource must be able to be used by any set of users in any context consistent with its overall goals.
- *Common*. The resource must present a common and consistent interface to all users, accessible by a standard mean. Thus *common* may be synonymous with the term *standard*.
- *Enabling*. The resource must provide the basis for any user or set of users to create, develop, and implement any applications, utilities, or services consistent with its goals.
- *Physical embodiment of an architecture.* The infrastructure is the physical expression of an underlying architecture. It expresses a world-view. This world view must be balanced with all the other elements of the infrastructure.
- *Enduring.* The resource must be capable of lasting for an extensive period of time. It must have the capability of changing incrementally and in an economically feasible fashion to meet the slight changes of the environment, but must be consistent with the worlds view. In addition it must change in a fashion that is transparent to the users.
- *Scale.* The resource can add any number of users or uses and can by its very nature expand in a structured manner in order to ensure consistent levels of service.
- *Economically sustainable.* The resource must have economic viability. It must meet the needs of both customers and providers of information products. It must provide for all elements of a distribution channel, bringing the product from the point of creation to the point of consumption. It must have all the elements of a food chain.

Infrastructures corresponding to this definition will to a large extent embody the three first infrastructure aspects mentioned above. Shareable and common in McGarty’s definition is literally the same as shared in ours, scale is close to openness. There is, however, one important difference between our definitions: McGarty's defines infrastructure as largely homogenous. There is no non-technical elements involved, and even the technology is required to homogeneous. This is connected to his notion of "physical embodiment of architecture." The requirement that the whole infrastructure should be based on one architecture is an expression of traditional engineering design ideals like uniformity, simplicity, consistency. (These ideals are also underlying McGarty's requirements about how an II should scale.) If it is a requirement that an II is based on one single architecture, it would be impossible to make a larger II by linking together two existing ones through a gateway if they are based upon different architectures (which different II regularly will be) even though doing this should be desirable as well as feasible. This requirement implies a form of homogeneity being in direct conflict with the essential heterogeneous nature of infrastructures. It implies a strategy for making *closed* systems. In this way, this definition is only open as far as one uniform coherent architecture is applicable and acceptable. It is the main argument of this book that this closed world thinking inherent in virtually all engineering activity (at least as presented in text books) is a main obstacle for successfully developing the II visions in documents like the US NII plan and the Bangemann report, an issue we now will turn to. The situation may be illustrated graphically as in figure xxx.

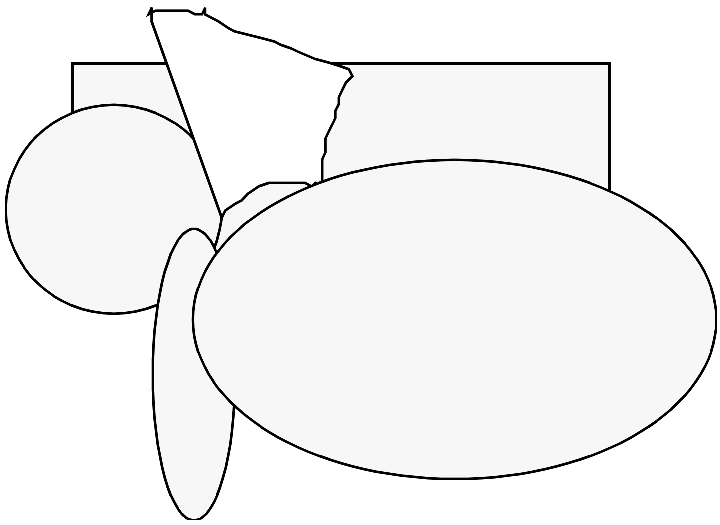

FIGURE 2. **The ecology of networks which give IIs their non-homogenous character**

The ecologies of infrastructures will be explored further in chapters 8 and 11.

Installed base

Building large infrastructures takes time. All elements are connected. As time passes, new requirements appear which the infrastructure has to adapt to. The whole infrastructure cannot be change instantly - the new has to be connected to the old. The new version must be designed in a way making the old and the new linked together and “interoperable” in one way or another. In this way the old - the installed base - heavily influence how the new can be designed.

> Aspect 6: Infrastructures develops through extending and improving the *installed base*.

The focus on infrastructure as “installed base” implies that infrastructures are considered as always already existing, they are NEVER developed from scratch. When “designing” a “new” infrastructure, it will always be integrated into or replacing a part of a later one. This has been the case in the building of all transport infrastructures: Every single road - even the first one if it make sense to speak about a such - has been built in this way; when air traffic infrastructures have been built, they have been tightly interwoven with road and railway networks - one needed these other infrastructures to travel between airports and the travels’ end points.Further, powerful telecommunications infrastructures are required to operate the air traffic infrastructure safely. Pure air traffic infrastructures can only be used for one part of a travel, and without infrastructures supporting the rest, isolated air traffic infrastructures would be useless.

In McGarty’s definition, his notion of enduring includes the requirement that infrastructures has to change incrementally to meet the changes of the environment. Such changes will be very, very modest if they always have to take place within the constrains of the original architecture. So we think it is fair to say that McGarty does not pay any attention to the importance of the installed base whatsoever.

The same kind of engineering ideals as found in McGarty’s definition played an important role in the development of the OSI suite of protocols. This effort failed - according to Einar Stefferud and Marshal T. Rose (ref.) due to the protocols’ “installed base hostility.” All protocols were designed without any consideration of how an OSI network should interoperate with others. It was assumed that OSI protocols would replace all others. However, as there was an installed base of other communication protocols, the OSI ones were never adopted as they could not (easily enough) be linked to these.

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

Star and Ruhleder (1996, p. 113) give a definition sharing some of the features of McGarty's and the NII one. However, in their definition Star and Ruhleder put more emphasis on the social relations constituting infrastructures. They characterize II by holding that it is "fundamentally and always a relation," and that "infrastructure emerge with the following dimensions:"

- *Embeddedness*. Infrastructure is "sunk" into, inside of, other structures, social arrangements and technologies;
- *Transparency.* Infrastructure is transparent in use, in the sense that it does not have to be reinvented each time or assembled for each task, but invisibly support those tasks;
- *Reach or scope*. This may be either spatial or temporal - infrastructure has reach beyond a single event or one-site practice;
- *Learned as part of membership.* The taken-for-grantedness of artifacts and organizational arrangements is a *sine qua non* of membership in a community of practice. Strangers and outsiders encounter infrastructure as a target object to be learned about. New participants acquire a naturalized familiarity with its objects as they become members;
- *Links with conventions of practice.* Infrastructure both shapes and is shaped by the conventions of a community of practice, e.g. the way that cycles of day-night work are affected by and affect electrical power rates and needs.
- *Embodiment of standards.* Modified by scope and often by conflicting conventions, infrastructure takes on transparency by plugging into other infrastructures and tools in a standardized fashion.
- *Built on an installed base.* Infrastructure does not grow *de novo*; it wrestles with the "inertia of the installed base" and inherits strengths and limitations from that base.
- *Becomes visible upon breakdown*. The normally invisible quality of working infrastructure becomes visible when it breaks.

The configuration of these dimensions forms "an infrastructure," which is without absolute boundary on a priori definition (ibid.).

Opposed to McGarty, this definition stresses the heterogeneous character of infrastructures as expressed in the notions of embeddedness as well as its socio-technical nature in forms by being linked to conventions of practice. Although the importance of the installed base is emphasised, Star and Ruhleder do not address design related issues or implications. Further, learned as part of membership and links with conventions of practice have important implications: IIs must support today's conventions at the same time as they must stimulate to their change if significant benefits is to be gained. These aspects of infrastructures also mean that although IIs are enabling and generic, they are not completely independent of their use. The interdependencies and possible co-evolution of IIs and conventions of practice will be explored in chapter yy.

Although having much in common, an interesting difference between the two definitions given by Star and Ruhleder and McGarty respectively is the fact that the first stresses the importance of the installed base and its inertia while the latter requires that an infrastructure must have the capability of being incrementally changed to meet new needs, and that this change must be transparent to users. The combination of - or tension between - these two elements is the core of IIs as seen in this article: understanding the nature of the installed base and how to cultivate it.

Seeing II and II development as "installed base cultivation" captures most aspects of (information) infrastructures as described above. IIs are larger and more complex systems, involving large numbers of independent actors as developers as well as users. This fact makes standards a crucial part of IIs. Further, IIs grow and develop over a long period of time, new parts are added to what exists and existing parts are replaced by improved ones. An II is built through extensions and improvements of what exists - never from scratch. It is open in the sense that any project, independent of how big it is, will just cover a part of an II. The rest exits already and will be developed by others being out of reach for the project and its control. What is developed by a defined closed activity will have to be hooked into an existing II. What exists has significant influence on the design of the new. In sum: IIs are developed through the cultivation of the installed base.

IIs and II development may to a large extent be described as specification and implementation of standards. Cultivating IIs may, accordingly, be considered as cultivating standards.

The role of the installed base in the development of infrastructures and standardized technologies will be discussed in chapter 9, while design strategies for infrastructures having the characteristics presented above will be the theme of chapters 10 and 11.

Sliding boundaries

The focus on IIs as open systems raises some questions - among these: Are the borders between open (IIs) and closed (I) systems absolute and predetermined in some sense? Is the crucial role played by the installed base a unique feature of infrastructure and systemic technologies or is it a more general one? Focusing on IIs as an open installed base means that IIs are never developed from scratch - they *always already exist*. If so - when and how is an II born?

Seeing IIs as an open installed base focuses on what makes them different from ISs. The latter may be described as closed in the sense that one project or defined activity may design and control the development of all its parts. Whether ISs should be considered as open or closed systems depends on the time frame chosen. Considering their life from initial design and development throughout the whole maintenance phase, no IS can properly be understood as closed (Kling and Iacono 1984).

Whether a system is open or closed is not always an a priori aspect of the system. It depends on the perspective chosen. Whether an open or closed perspective is most appropriate in a specific situation is often not an easy question. And whether choosing an open or closed perspective deals with our basic conception of reality. Each perspective is linked to a wide range of other perspectives, theories, tools and techniques, etc. The complexity and difficulties related to this issue can be illustrated by "discussions" and disagreements between Popper and Feyerabend about the nature of science and proper scientific work, and fights between positions like positivism and dialectics. This issue has also been involved in discussions about the approaches followed by OSI and Internet standardization efforts. Einar Stefferud (1994) summarizes the "religious war" (Drake 1993) as a discussions between two competing and incommensurable paradigms, one being an open *internetworking* paradigm, the other being a closed *networking* one.

ISs and IIs are not totally disjunct - there is a border area. An IS, e.g. an interorganizational system (IOS) or distributed IS (DIS), may "travel" through this area and change and become an II. IIs are born when

1. new and independent actors become involved in the development of an IOS or DIS so that the development is not controlled by one actor any more;
2. existing IOSs are linked together and the development of the linked networks is not controlled by one actor;
3. an IOS/DIS may be considered an II when the goal is that it shall grow an become an II (or part of) in the future and this is an important design objective.

Where to draw the borderline between for instance an IOS and an II also depends on which aspect of an infrastructure definition is emphasized. Taking flight booking systems as an example, they may be considered infrastructure as they are large, complex and shared by a large user community. However, they are specialized applications rather than generic, enabling substructures.

Installed base is not a unique II feature. In general the issue is the importance of history, the history's imprint on the future. History is everywhere, in IS design as well as II cultivation. However, the issue needs to be dealt with in another sense when building II. In the social sciences, similar issues is studied under the label "new institutionalism" (Scott 1995, North 1990, Powell and DiMaggio 1991, March and Olsen 1989), and in philosophy by Heidegger in his Being and Time.

State of the art

The growing interest for information infrastructure has produced a rich variety of studies and analyses of information infrastructures. There do not, surprisingly enough, exist many studies about the Internet which try to spell out in some detail how the design process actually takes place. There exist several overviews of the historical development of Internet but they contain little or no evidence of how or why various design decisions came about (see for instance Hauben and Hauben 1006; LoC 1996). Abbate (1994) represents an exception. Here the underlying design visions of two competing alternatives for networking, namely IP and the one developed within the telecommunication community (called X.25 by the CCITT[1]), are uncovered. Hanseth, Monteiro and Hatling (1996) discuss the structure of the tension between change and stability in information infrastructures with illustrations from Internet and the Open Systems Interconnection (OSI) of the International Standardisation Organisation (ISO). Hanseth (1996) analyses the nature of the installed base through illustrations of a variety of cases, including Internet. A highly relevant area of research regarding Internet is to unwrap the design culture within Internet. This, however, seems to be a completely neglected area. The few studies related to cultural aspects of Internet focus on others than the designers, for instance, Turkle (1995) on MUD users, Baym (1995) on Usenet users.

1. CCITT is the international standardisation body for standards within telecommunication.

Our approach to the study of standardisation resembles those applied by Marc Berg, Geoffrey Bowker, Leigh Star and Stefan Timmermans (Bowker and Star 1994; Bowker, Timmermans, and Star 1995; Star and Ruhleder 1996, Timmermans and Berg 1997). In their studies of classification schemes and infrastructures Bowker and Star identify a number of issues which are closely related to those we are focusing on. In a study of the evolution of the classification of diseases maintained by the World Health Organisation, they illustrate how coding and classification — essential tasks in the establishment of an information infrastructure — is anything but neutral. Interests are inscribed into coding schemes (Bowker and Star 1994). Bowker, Timmermans, and Star (1995) study how some aspects of work is made more visible than other by inscribing them into a classification scheme. Star and Ruhleder (1996) discuss key characteristics of infrastructure based on a study of the introduction and use of an information infrastructure.

Within the field of social studies of technology, there are some contributions relevant to a study of information infrastructure standardisation. Some focus on conceptual issues, for instance, the work by Fujimura (1992) on standardising procedures and interpretations across geographical distances. Others explore empirically how universal standards are appropriated to local contexts (Berg 1995) or how the interplay between stability and change is played out (Hanseth, Monteiro and Hatling 1996). Similarly, Jewett and Kling (1991) develop a notion of infrastructure which is to capture the many hidden resources which need to be mobilised to get an information system to actually be used.

Other studies of the standardisation process of information infrastructure tend to focus issues rather different from ours and those mentioned above. These include the economical significance of standards (David 1987; OECD 1991), technical challenges (Rose 1992), the use of information infrastructures (Ciborra 1992), the political nature of standardisation bodies (Webster 1995) or cultural differences (Trauth, Derksen and Mevissen 1993). The predominant view on information infrastructure standardisation is that it is straightforward. An exception to this is (Williams 1997). Standardisation is commonly portrayed either as (i) a fairly unproblematic application of mainstream techniques for solving the technical difficulties of software engineering or (ii) it is simply glossed over or taken as given (Ciborra 1992; David 1987; OECD 1991). Those involved in the design of information infrastructure have so far been very close to (i). These studies shed little light on the socio-technical complexity of establishing an information infrastructure.

Lehr (1992) points to the bureaucratic and procedural differences in the way standardisation bodies organise their work. These are argued to play an important role for the outcome, namely the standards. For instance, the OSI effort represents a clear-cut alternative to the evolutionary approach underlying an emphasis on transition strategies. OSI is designed monolithically from scratch, that is, with a total disregard for existing information infrastructures, the installed base. It has been fiercely criticised for exactly this (Hanseth, Monteiro and Hatling 1996; Rose 1992).

The literature on large technical systems is illuminating in describing how infrastructures are established but tend to bypass how to facilitate changes (Summerton 1994). A particularly relevant contribution is (Hughes 1983) which gives an historical account of the electrification of the Western world around the turn of the century. Hughes' work is important but it does not address the dilemmas of scaling explicitly. It provides, however, a rich empirical material containing lots of illustrations of transition strategies, the role of the installed base and gateway technologies, etc.

Recently, there has been attention to development strategies suitable for information infrastructures (Kahin and Abbate 1995). These strategies do not deal with scaling but address issues such as the role of government intervention and industrial consortia.

Grindley (1995) argues for the importance of the installed base of products. This emphasises the need for products to be backwards compatible, that is, that they interoperate with earlier versions of the product. In other words, this protects the installed base of earlier versions of the product. Backward compatibility plays the same role for products as transition strategies for information infrastructures (Hinden 1996).

CHAPTER 4

Standards and standardization processes

Introduction

The world is full of standards. Standards regulate, simplify and make possible an extensive division of labour which should be recognized as a necessary basis for far-reaching modernization processes (REFs noe STS).

Also in the world of computers there is a rich variety of standards. The conventional wisdom, however, is that standards are either simple and straightforward to define or purely technical (REFs). One, if not *the*, key theme running right through this book is that the development of an information infrastructure, necessarily including the standards, should instead be recognized as a highly complex socio-technical negotiation process. It is a pressing need to develop our understanding of how “social, economic, political and technical institutions (...) interact in the overall design of electronic communication systems” (Hawkins 1996, p. 158). There is accordingly a need to classify and conceptualize to grasp the role of standards in the development of information infrastructures.

This chapter first outlines the basic argument for standards for communication technology. Strictly speaking, “Communication systems cannot function without standards” (Hawkins 1996, p. 157). We then provide a taxonomy of different types of standards. Standards for information infrastructures are worked out within quite distinct institutional frameworks. We describe the most influential, international institutions aiming at open information infrastructures. Lastly, we briefly review the literature on standardization.

The economy of scale argument

Standardized technology abound and make perfectly good sense. It simplifies otherwise complicated choices, enables large scale integration, and it is the basis for a division of labour. The standardization of the design of cars created such a division of labour between car manufacturers and suppliers of standardized parts ranging from roller bearings and lamps to complete motors (når? ref). For communication technology there is in addition a very influential, brute force argument. It is simple and is typically illustrated by the following figure.

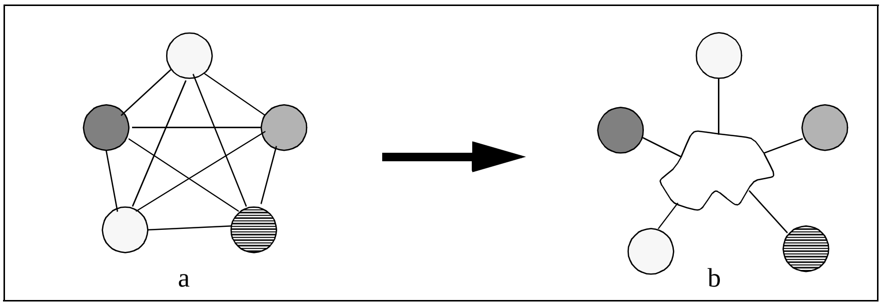

FIGURE 3. The number of different links as a function of the number of nodes.

The figure shows how the number of communication connections, or protocols, (the edges) rapidly increases as the number of communicating partners (the nodes) is rising. In the left hand case, every pair of communicating partners need a protocol to communicate. With 4 partners the required number of protocols is 6. The number of protocols increase rapidly. With 5 partners, the number of required protocols is 10, with 6 partners 15, etc. The number of protocols is given by the formula n(n-1)/2, where n is the number of nodes. In the right hand situation, all communication is based on one single shared, standardized protocol. What is needed is establishing a link between each partner (node) and the standard. As a consequence, the number of required links to be established and maintained increases linearly with the number of partners — given the existence of a shared standard. With 4 partners, 4 links are required and with 5 partners, 5 links are required, etc. Already with a relative small community of communicating partners, the only feasible strategy is to use a shared, standardized protocol, an Esperanto language which everyone reads and writes. A solution with pairwise protocols among the communicating partners simply does not scale, it works in principle but not in practise. Larger communication networks is simply impossible to build and manage if not based on standards.

As mentioned in the previous chapter, we consider standards not only important from a practical and economic perspective, but also an essential and necessary constituting element. If an "infrastructure" is built only on the bases of bilateral arrangements, it is no real infrastructure. It is then just a collection of separate independent connections.

The brute force argument sketched above makes a lot of sense. The ideal of establishing a shared standard is easy to understand and support. The problem, however, is how to decide which areas should be covered by one standard, how different standards should relate to each other and how to change them as their environment changes, i.e how to pragmatically balance the idealized picture of everybody sharing the same standard against the messy, heterogeneous and irreversible character of an information infrastructure. This act of balancing — as opposed to dogmatically insisting on establishing one, shared standard — is very close to the heart of the design of an information infrastructure. It is an issue we explore in greater in the subsequent chapters.

Types of standards

Standards abound. David and Greenstein (1990, p. 4) distinguish among three kinds of standards: reference, minimum quality and compatibility standards. II standards belong to the last category, that is, standards which ensure that one component may successfully be incorporated into a larger system given an adherence to the interface specification of the standard (ibid., p. 4). One may also classify standards according to the processes whereby they emerge. A distinction is often made between formal, de facto and de jure standards. Formal standards are worked out by standardisation bodies. Both OSI and Internet are formal according to such a classification.[1] De facto standards are technologies standardised through market mechanisms, and de jure standards are imposed by law.

De facto standards are often developed by industrial consortia or vendors. Examples of such standards are the W3 consortium currently developing a new version of the HTML format for WorldWideWeb, IBM´s SNA protocol, CORBA developing a common object oriented repository for distributed computing, X/Open developing a new version of Unix and the Health Level 7[2] standard for health care communication. Some of these consortia operate independently of the international standardisation bodies, others align their activities more closely. For instance, the W3 consortium is independent of, but closely linked to, the standardisation process of the IETF (see further below).

Internet standards

The Internet is built of components implementing standardized communication protocols. Among these are the well known ones such as TCP, IP, SMTP (email), HTTP (World Wide Web), FTP (file transfer) and TELNET (remote login). But Internet includes many more standards - in June 1997 there were in fact 569 officially registered Internet standards (RFC 2200). These standards are split into different categories.

Maturity levels and status

There are two independent categorization of protocols. The first is the “maturity level” which in the Internet terminology is called the state of standardization. The state of a protocol is either “standard”, “draft standard”, “proposed standard”,

1. This is the source of some controversy. Some prefer to only regard OSI as “formal” due to properties of the standardisation process described later. This disagreement is peripheral to our endeavour and is not be pursued in this book.

2. Health Level 7 is a standard worked out by an ad-hoc formation of a group of smaller vendors in the United States, later on being affiliated to American National Standards Institute, ANSI (see url http://www.mcis.duke.edu/standards/HL7/hl7.htm).

“experimental”, “informational” or “historic”. The second categorization is the “requirement level” or status of a protocol. The state is either “required”, “recommended”, “elective”, “limited use”, or “not recommended”.

When a protocol is advanced to proposed standard or draft standard, it is labelled with a current status.

In the Internet terminology computers attached to or otherwise a part of the network is called a “system.” There are two kinds of systems - hosts and gateways. Some protocols are particular to hosts and some to gateways; a few protocols are used in both.. It should be clear from the context of the particular protocol which types of systems are intended.

Protocol states are defined as follows:

- *Standard Protocol:* The IESG (Internet Engineering Steering Group) has established this as an official standard protocol for the Internet. These protocols are assigned STD numbers (see RFC- 1311). These are separated into two groups: (1) IP protocol and above, protocols that apply to the whole Internet; and (2) network-specific protocols, generally specifications of how to do IP on particular types of networks.
- *Draft Standard Protocol:* The IESG is actively considering this protocol as a possible Standard Protocol. Substantial and widespread testing and comment are desired. Comments and test results should be submitted to the IESG. There is a possibility that changes will be made in a Draft Standard Protocol before it becomes a Standard Protocol.
- *Proposed Standard Protocol:* These are protocol proposals that may be considered by the IESG for standardization in the future. Implementation and testing by several groups is desirable. Revision of the protocol specification is likely.
- *Experimental Protocol:* A system should not implement an experimental protocol unless it is participating in the experiment and has coordinated its use of the protocol with the developer of the protocol.
- *Informational Protocol:* Protocols developed by other standard organizations, or vendors, or that are for other reasons outside the purview of the IESG, may be published as RFCs for the convenience of the Internet community as informational protocols.
- *Historic Protocol:* These are protocols that are unlikely to ever become standards in the Internet either because they have been superseded by later developments or due to lack of interest.

Typically, experimental protocols are those that are developed as part of an ongoing research project not related to an operational service offering. While they may be proposed as a service protocol at a later stage, and thus become proposed standard, draft standard, and then standard protocols, the designation of a protocol as experimental may sometimes be meant to suggest that the protocol, although perhaps mature, is not intended for operational use.

Protocol status is defined as follows:

- *Required Protocol:* A system must implement the required protocols.
- *Recommended Protocol:* A system should implement the recommended protocols.
- *Elective Protocol:* A system may or may not implement an elective protocol. The general notion is that if you are going to do something like this, you must do exactly this. There may be several elective protocols in a general area, for example, there are several electronic mail protocols, and several routing protocols.
- *Limited Use Protocol:* These protocols are for use in limited circumstances. This may be because of their experimental state, specialized nature, limited functionality, or historic state.
- *Not Recommended Protocol:* These protocols are not recommended for general use. This may be because of their limited functionality, specialized nature, or experimental or historic state.

TABLE 1. Internet standards - numbers and growth

	94/7[a]	**Added**[b]	**97/6**[c]	**New**[d]	**Removed**[e]
Standard	58	2	60	7	5
Draft	21	28	49	45	17
Proposed	161	99	260	150	51
Experimental	53	45	98	55	10
Informational	23	34	57	54	20
Historic	35	10	45	10	0
Totally	351	218	569	321	103

a. Number of Internet standards in July 1994 (RFC 161o).

b. Increase in the number of Internet standards from July 1994 to June 1997 (RFCs 1720, 1780, 1800, 1880, 1920, 2000, 2200).
c. Number of Internet standards in June 1997 (RFC 2200).
d. Total number of new Internet standards from July 1994 to June 1997 (RFCs 1720, 1780, 1800, 1880, 1920, 2000, 2200).
e. The number of Internet standards deleted from the official list from July 1994 to June 1997 (ibid.).

Among the 569 registered Internet standards there are 60 standard protocols,[3] 49 draft standards, 260 proposed, 98 experimental, 57 informational and 45 historic. The growth in standards from July 1994 is, as illustrated in Table 1, quite significant - about 62%. However, the growth has been rather uneven among the various standards categories - 3.5% growth in full standards, 133% in draft and 60% in proposed standards. Looking at the numbers in Table 1, we see that the number of proposed and experimental standards is growing rapidly. Lots of new standards are popping up, while close to none is advancing to the top level. An important explanation for this is the increasing growth of the Internet, both in numbers of computers (and users) connected and in the complexity of the technology. Within the environment constituted by this complex unstable environment, developing stable and accepted standards gets increasingly more difficult (Alvestrand 1996). This points to the more general question of whether the organisation of Internet standardisation which so far has proven so successful, has reached its limit, i.e. that there is a need for a more formal, more ISO like organisation (see also earlier remarks in chapter 2). Alternatively, the size the Internet is approaching represents a limit for big far that kind of networks can grow.

Types of Internet standards

The Internet standards specifies lots of different kinds of communication (sub-) technologies. We will here mention some of them just to give a flavour of what all these protocols are all about. Interested readers should consult the truly vast electronic archive the Internet community keeps of its reports and discussion (see the list cited in the section on methodological issues in chapter 2).[4]

3. There are 53 Internet standards assigned a number. We have here counted the number of RFCs included in the official list of standards.

4. A good place to start looking for RFCs and information about Internet standards is http://ds.internic.net.

First of all there are *protocol* specifications like TCP, IP, SMTP, HTTP, RTP (Transport protocol for Real Time Applications), FTP, etc. These are the standards most often associated with the Internet.

In addition, there are lots of *auxiliary protocols*, protocols offering services to the others. These include PPP (Point-to-Point Protocol), SNMP (for network management), Echo protocol, DNS (Domain Name System, a distribute data base system mapping a symbolic Internet address (like diagnostix.ifi.uio.no) to a numerical one (like 129.240.68.33)), tools for DNS debugging, Time Server Protocol, etc. There are further lots of standards defining security systems (Kerberos authentication service, encryption control protocol, MIME object security services, Signed and encrytped MIME, etc.).

The Internet runs across many different physical networks, accordingly there are standards defining how to *implement one protocol on top of these networks*, typically IP on ARPANET, Wideband Network, Ethernet Networks, IEEE 802, transmission of IP traffic over serial lines, NETBIOS and FDDI. There is also a significant number of standards defining *gateways* between protocols, like FTP - FTAM Gateway Specifications and X.400 - MIME Body Equivalencies. One group of standards defines how to construct various *identifiers* (including addresses) like IP addresses and URL specifications.

SNMP is the Internet network management protocol, defining the general rules for management of any part of the Internet. However, to set up a management system, additional information about the specific networks and protocols is required. Accordingly, there is one standard, SMI, defining how this information should be specified as well as standards (MIBs, Management Information Bases) defining how to manage specific parts of the Internet using for instance ATM, Ethernet, IP, TCP, UDP, DECNET and X.500.

Lastly, there are standardized data formats like MAIL (Format of Electronic Mail Messages), MIME (Multipurpose Internet Mail Extensions) and MIME extensions (MIME Media Types, MIME message header extensions for non-ASCII), HTML, SGML Media Types, Using Unicode with MIME, NETFAX (File format for exchange of images), MIME encapsulation of Macintosh files and Serial number arithmetic.

Standards for health care information infrastructures

We give a brief outline of the different II standards related to the health care sector. A comprehensive overview of various international standardization efforts can be found in (De Moor, McDonald and van Goor 1993).

CEN TC/251

In chapter 2 we presented CEN TC/251 as the standardisation body most widely accepted as the authoritative one. We will in this section give a brief overview of what kinds of standards this body is defining. CEN/TC 251 has split its work into seven different subfields which are called:[5]

1. Healthcare Terminology, Semantics and Knowledge Bases
2. Healthcare Information Modelling and Medical Records
3. Healthcare Communications and Messages
4. Medical Imaging and Multimedia
5. Medical Device Communication in Integrated Healthcare
6. Healthcare Security and Privacy, Quality and Safety
7. Intermittently Connected Devices

Within these areas standards of very different types are defined. We will here illustrate this with the standards defined by xx 1996 within four areas. Within "Healthcare Terminology, Semantics and Knowledge Bases," standards with the following titles were defined:

- Structure for Nomenclature, Classification and Coding of Properties in Clinical Laboratory Sciences
- Structure for Classification and Coding of Surgical Procedures
- Categorical Structures of Systems of Concepts - Model for the Representation of Semantics
- Medicinal Product Identification Standard
- Structure of Concept Systems - Medical Devices

5. Information about the work of CEN TC/251is found at http://

Within the second area, “Healthcare Information Modelling and Medical Records,” standards were defined with the following titles:

- Medical Informatics - Vocabulary,
- Healthcare Information Systems Framework,
- Electronic Healthcare Records Architecture
- Standard Architecture for Healthcare Information Systems.

Within “Healthcare Communications and Messages” the standards defined were

- Procedures for the Registration of Coding Schemes related to Healthcare
- Messages for Exchange of Laboratory Information
- Recording Data Sets used for Information Interchange in Healthcare
- Request and Report Messages for Diagnostic Services Departments
- Messages for Exchange of Healthcare Administrative Information
- Messages for Patient Referral and Discharge
- Methodology for the Development of Healthcare Messages

Within “Medical Imaging and Multimedia” the following standards were defined:

- Medical Imaging and Related Data Interchange Format Standards
- Medical Image Management Standard
- Medical Data Interchange: HIS/RIS-PACS and HIS/RIS - Modality Interfaces
- Media Interchange in Medical Imagining Communications

Each of these standards are also given a status similar to those used in the Internet community reflecting its maturity level (pre-standard, standard).

The standards cover a wide range of different health care related phenomena. They are also related to very different aspects of information infrastructure development and use. Several specify “messages,” i.e. structure of information to be exchanged. These standards specify also the institutions (or partners) the information should be exchanged between and when (for instance that a lab report should be sent from the lab to the ordering unit when the ordered analysis are finished). For these standards it is also defined by a separate group of standards the formats and protocols to be used for exchanging them. Another group of standards defines the semantics of important data fields in terms of nomenclatures and classification systems to be used. Further, there are standards defining the overall architectures of health care information systems and infrastructures, and even methodologies for developing various types of standards.

While CEN standards are those having been considered the most important in Europe, there are lots of others. Some of these are overlapping, and in some cases one standard defined by one body is simply adopted by another. For instance, the DICOM standard for medical images developed by ACR/NEMA has a broad acceptance in this area and is accordingly more or less adopted by CEN as it is. Similarly, the EDIFACT messages defined by CEN is also given official status within the EDIFACT bodies as well.

EDI

EDI denotes electronic exchange of form-like information between different organization. It is often used to transfer electronic equivalents of already standardized paper forms. Examples are orders, invoices, customs declaration documents, bank (payment) transactions, etc. Within health care there is a vast range of such forms, including lab orders and reports, admission and discharge letters and drug prescriptions.

Within EDI, the EDIFACT format is the most widely accepted standard (among standardization bodies, may be not among users). Accordingly, EDIFACT has been chosen as a basis for exchange of form-like information in health care as well.

In the EDIFACT community, there are three forms of standards: messages, segments and data elements. The electronic equivalent of a paper form is a message. A message is, in EDIFACT terms, composed of segments and segment groups, where a segment group is (recursively) defined by segment groups/and or segments. A segment is composed of data elements - single or composed. The latter being composed by a number of single ones.

EDIFACT messages are defined internationally. Such international messages are often accompanied by specification of national or sectorwise subsets as well as “exchange agreements” between pairs of communicating partners. Such specifications define in more detail how a general message should be used within a region or sector or between specific partners.

Standardisation processes and strategies

Information infrastructures, like many other kinds of large technical systems (Hughes 1987), are standardised by formal, quasi-democratic, international standardisation bodies (Lehr 1992). These standardisation bodies have to follow predefined procedures and rules regulating the status, organisation and process of developing standards. In recognition of the limits of both market forces and hierarchical control, formal standardisation is a key strategy for developing an information infrastructure (OECD 1991).

Different standardization institutions organise the process of standardisation quite differently along several important dimensions, including the way participation in the process is regulated, how voting procedures are organised, the requirements proposed standards have to meet at different stages in the process, the manner information about ongoing standardisation is made public, and the bureaucratic arrangements of how work on one, specific standard is aligned with other efforts, etc.

Standardization processes has only recently become a topic for research and debate. Branscomb and Kahin (1995) discuss three possible models for NII standards development, quite close to David's three categories mentioned above, which they have given the following names:

1. The Applications Model: Intense Competition and Ad Hoc Consortia.
2. The Internet Model: A Cooperative Platform for Fostering Competition.
3. The Telecommunications Model: From National-Level Management to Open Competition

We will now present OSI and EDIFACT as an example of the telecommunications model and the Internet model. We will also present standardization processes and strategies adopted in health care, which in fact combines all models to some extent.

OSI

OSI is short for the ISO Open Systems Interconnection (OSI) reference model. OSI was worked out by the International Standardization Organization (ISO) in the early 80s. ISO is a voluntary, non-treaty organization that was founded just after World War II and produce international standards of a wide range of types. The members of ISO are national standardization bureaus rather than individuals or organizations. Members include national standardization organisations like ANSI (from the US), BSI (from Great Britain), AFNOR (from France) and DIN (from Germany) (Tanenbaum 1989).

ISO is divided into a number of technical committees according to whether the focus is on specification of nuts and bolts for construction work, paint quality or computer software. There are several hundreds technical committees within ISO. The “real” work is done by the several thousand non-paid volunteers in the working groups. A number of working groups belong to a subcommittees which, again, belong to a technical committee. To refer to the members of the working groups as non-paid volunteers merely implies that they are not paid or employed by ISO as such. They are typically employed by large vendors, consumer organizations or governmental institutions. Historically, vendors have dominated standards committees (Jakobs 1998; Lehr 1992).

The development of OSI protocols follow (in formal terms) democratic procedures with representative participation under the supervision of the ISO (Lehr 1992). The standardization process aims at achieving as broad consensus as possible. Voting is based on representative voting, that is, that each member (representatives of national bureaus of standardization) is given a given weight.

An ISO standard passes through certain stages. It starts as a draft proposal and is worked out by a working group in response to one representative’s suggestion. The draft proposal is circulated around for six months and may be criticized. Given that a substantial majority is favourable, criticism is incorporated to produce a revised document called a draft international standard. This is circulated for both comments and voting. This then is fed into the final document, an international standard, which get approved and published. When faced with controversial issues, the process may back-track in order to work out compromises that mobilize sufficient support in the voting. In this way, as was the case with OSI, the standardization process may stretch over several years.

OSI protocols are developed by first reaching a consensus about a specification of the protocol. The protocol specifications are assumed to be implemented as software products by vendors. The implementation is independent of the standardisation process. Because of the formal and political status of OSI protocols, most Western governments have decided that II in the public sector should be based on OSI protocols.

The implementation and diffusion of OSI protocols have not proceeded as anticipated by those involved in the standardisation processes. One of the main reasons is that they have been developed by large groups of people who have been specifying the protocols without any required implementation and without considering compatibility with non-OSI protocols (Rose 1992). This results in very complex protocols and serious unforeseen problems. The protocols cannot run alongside other networks, only within closed OSI environments. The protocols are big, complex and ambiguous, making them very difficult to implement in compatible ways by different vendors. The definition of profiles mentioned earlier is an attempt to deal with this problem.

EDIFACT

EDIFACT, short for Electronic Data Interchange for Administration, Commerce and Transport, is a standardization body within the United Nation. This has not always been the case. EDIFACT has during the last couple of decades transformed dramatically both in content and institutionally. It started in the early days, as an informal body of about 30 people world-wide (Graham et al. 1996, p. 9). Since then it has grown to a huge, global bureaucracy involving several hundred participants. The small group of people who established EDIFACT *chose* the United Nation as their overarching organization because they expected the perceived neutrality of the United Nation to contribute favourably to the diffusion of EDIFACT messages (ibid., p. 9). The representation in EDIFACT is, as is usual within the United Nations, through national governments rather than the national standards bodies as would have been the case if EDIFACT had chosen to align with ISO instead.

During 1987, EDIFACT reorganized into three geographically defined units, North America, Western Europe and East Europe. This has later been extended to cover the areas Australia/ New Zealand and Japan/ Singapore. These areas are subsequently required to coordinate their activity. Although these areas in conjunction cover a significant part of the world, the vast majority of the work has taken part within the Western European EDIFACT board (ibid., p. 9). This is due to the close alignment with the TEDIS programme of the Commission of the European Community. In 1988, the EDIFACT syntax rules were recognized by the ISO (LIASON???XXXX).

EDIFACT messages pass through different phases before reaching the status as a proper standard. A draft is first circulated to the secretariat and all other regional secretariat to be assessed purely technically before being classified as status level 0. Moving up to status 1 requires a consensus from all the geographical areas. After status 1 is achieved, the proposal has to be delayed for at least a year to allow implementations and assessments of use. If and when it reaches status 2, it has become an United Nation standard and is published.

The EDIFACT organization also specifies rules for design of messages. For instance, if an existing message can be used, that one should be adopted rather than designing a new one. In the same way the rules specify that the existing segments and data elements should be reused as far as possible.

Internet

As illustrated in chapter 2, the organization of the development of the Internet has changed several times throughout its history. The organizational structure has changed from initially that of a research project into one being very close to the standard standardization body. Along the change in organizational structure have the rules also changed.

Internet is open to participation for anyone interested but without ensuring representative participation.[6] The development process of Internet protocols follows a pattern different from that of OSI (RFC 1994; Rose 1992). A protocol will normally be expected to remain in a temporary state for several months (minimum six months for proposed standard, minimum four months for draft standard). A protocol may be in a long term state for many years.

A protocol may enter the standards track only on the recommendation of the IESG; and may move from one state to another along the track only on the recommendation of the IESG. That is, it takes action by the IESG to either start a protocol on the track or to move it along.

Generally, as the protocol enters the standards track a decision is made as to the eventual STATUS, requirement level or applicability (elective, recommended, or required) the protocol will have, although a somewhat less stringent current status may be assigned, and it then is placed in the proposed standard STATE with that status. So the initial placement of a protocol is into the state of proposed protocol. At any time the STATUS decision may be revisited.

The transition from proposed standard to draft standard can only be by action of the IESG and only after the protocol has been proposed standard for at least six months.

The transition from draft standard to standard can only be by action of the IESG and only after the protocol has been draft standard for at least four months.

Occasionally, the decision may be that the protocol is not ready for standardization and will be assigned to the experimental state. This is off the standards track, and the protocol may be resubmitted to enter the standards track after further work. There are other paths into the experimental and historic states that do not involve IESG action.

Sometimes one protocol is replaced by another and thus becomes historic, or it may happen that a protocol on the standards track is in a sense overtaken by another protocol (or other events) and becomes historic.

Standards develop through phases which explicitly aim at interleaving the development of the standard with practical use and evaluation (RFC 1994, 5). During the first phase (a Proposed Standard), known design problems should be resolved but no practical use is required. In the second phase (a Draft Standard), at least two independent implementations need to be developed and evaluated before it may pass on to the final phase, that is, to be certified as a full Internet Standard. This process is intended to ensure that several features are improved, the protocols are

6. The term "Internet" may denote either (i) the set of standards which facilitate the technology, (ii) the social and bureaucratic procedures which govern the process of developing the standards or (iii) the physical network itself (Krol 1992; RFC 1994). This might create some confusion because a version of Internet in the sense of (i) or (iii) has existed for many years whereas (ii) is still at work. We employ the term in the sense of (ii) in this context. To spell out the formal organisation of Internet in slightly more detail (RFC 1994), anyone with access to Internet (that is, in the sense of (iii)!) may participate in any of the task forces (called IETF) which are dynamically established and dismantled to address technical issues. IETF nominates candidates to both Internet Advisory Board (IAB, responsible for the overall architecture) and Internet Engineering Steering Group (IESG, responsible for the management and approval of the standards). The IAB and IESG issue all the official reports which bear the name "Requests For Comments" (RFC). This archive was established along with the conception of the Internet some 25 years ago. It contains close to 2000 documents including: all the formal, proposed, draft and experimental standards together with a description of their intended use. The RFCs also record a substantial part of the technical controversies as played out within working groups established by the IETF or independent comments. Minutes from working group meetings are sometimes published as RFCs. In short, the RFCs constitute a rich archive which shed light on the historic and present controversies surrounding Internet. It seems to be a rather neglected source for information and accordingly an ideal subject matter for an informed STS project providing us with the social construction of Internet. It is an electronic achieve which may be reached by World-WideWeb using http://ds.internic.net.

lean and simple, and they are compatible with the already installed base of networks. Internet standards are to function in a multi-vendor environment, that is, achieve "interoperability by multiple independent parties" (ibid., p. 5).

A key source for identifying design principles shared by the vast majority of the Internet community, is the ones embedded in the procedural arrangements for developing Internet standards. The standards pass through three phases which explicitly aim at interleaving the development of the standard with practical use and evaluation:

> These procedures are explicitly aimed at recognizing and adopting generally-accepted practices. Thus, a candidate specification is implemented and tested for correct operation and interoperability by multiple independent parties and utilized in increasingly demanding environments, before it can be adopted as an Internet Standard.
>
> *(RFC 1994, p. 5)*

The Internet community consists, in principle, of everybody with access to Internet (in the sense of (ii) above) (RFC 1994). Participation in the e-mail discussions, either general ones or those devoted to specific topics, is open to anyone who submits an e-mail request in the way specified (see http://www.ietf.org). The Internet community may participate in the three yearly meetings of the Internet Engineering Task Force (IETF). The IETF dynamically decides to establish and dismantle working groups devoted to specific topics. These working groups do much of the actual work of developing proposals. At the IETF meetings design issues are debated. It is furthermore possible to organise informal forums called BOFs ("birds of feather") at these meetings.

IETF nominates candidates to both the 13 members of the Internet Advisory Board (IAB) and the 10 members of the Internet Engineering Steering Group (IESG). The IETF, the IESG and the IAB constitute the core institutions for the design of Internet. Their members are part-time volunteers. In principle, they have distinct roles: the IETF is responsible for actually working out the proposals, the IESG for managing the standardisation process and the IAB for the overall architecture of the Internet together with the editorial management for the report series within Internet called Requests For Comments (RFC). In practise, however, the "boundaries of the proper role for the IETF, the IESG and the IAB are somewhat fuzzy" as the current chair of the IAB admits (Carpenter 1996). It has proven particularly difficult, as vividly illustrated in the case further below, to negotiate how the IAB should exercise its role and extend advice to the IESG and the IETF about the overall architecture of the Internet protocols.

Standardization of health care IIs

The organization of CEN standardization work

CEN is the European branch of ISO follows the same rules and organizational principles as ISO in the definition of OSI protocols. Standards and work programmes are approved in meetings where each European country has a fixed number of votes. As presented above, the work is organized in eight so-called work groups, from WG1 up to WG8. Each group is responsible for one area. The tasks demanding "real work", for instance the development of a proposal for a standardized message, are carried out in project teams. More than 1500 experts have been involved (De Moor 1993).

In areas where EDIFACT is used, the definition of the EDIFACT message is delegated to the EDIFACT standardization body, WEEB (Western European EDIFACT board), which has established a so-called "message design group," MD9, for the health sector. They have a liaison agreement regulating their cooperation. This alignment is furthermore strengthened by the fact that a number of the members of WG3 within CEN TC 251 are also members of WEEB MD9. As of July 1995, the secretary of WE/EB was moved to CEN.

On the national level, standardization work is organized to mirror that on the European. Typical work tasks include specifying national requirements to a European standard, validation of proposed European standard messages and appropriating European standardized messages according to national needs.

CEN/TC 251 is following the ISO/OSI model.

Industrial consortia

To the best of our knowledge, there is no standardization efforts related to health care that can be said to follow the Internet model. Industrial consortia, on the other hand, play a significant role. The earlier mentioned HL-7 standardization is organized by one such. This organization has, or at least had in its earlier days, some rules (may be informal agreement) saying that the larger companies were not allowed to participate.

Within the field of medical imaging, the standardization work organized by ACR/NEMA (American College of Radiology/National Electronic Manufacturers' Association), developing the DICOM standard, seems by far to be the most influential. ACR/NEMA is an organization that could be seen as a rather traditional standardization body. However, it is in this field operating as an industrial consortium and the work is dominated by large companies like General Electric, Siemens and Phillips.

Standards engineering

Expositions like the one in this chapter of the contents and operations of the various standardisation bodies involved in information infrastructure development might easily become overwhelming: the number and complexity of abbreviations, institutions and arrangements is impressive (or alarming). The aim of this book deals with the design — broadly conceived — of information infrastructure and hence needs to come to grips with the practical and institutional setting of standardisation. Still, it is fruitful to underscore a couple of themes of particular relevance to us that might otherwise drown in the many details surrounding standardisation.

Specification or prototyping

The three models presented by Branscomb and Kahin (REF) categorize standardization processes according to primarily according to organizational and governmental principles. We will here draw attention to important differences between the OSI and Internet processes seen as technological design strategies. Even if never made explicit, we believe that the two approaches followed by, on the one hand, OSI (and EDIFACT and CEN) and, on the other hand, Internet could be presented as two archetypical approaches to the development of II. To explain this, we attempt to make explicit some of the underlying assumptions and beliefs.

The principal, underlying assumption of OSI's approach is that standards should be developed in much the same way as traditional software engineering, namely by first specifying the systems design, then implementing it as software products and finally put it into use (Pressman 1992). Technical considerations dominate. As for traditional software engineering (Pressman 1992, 771), OSI relies on a simplistic, linear model of technological diffusion, and in this case, for the adoption of formal standards. The standardisation of Internet protocols are based on different assumptions. The process is close to an approach to software development much less widely applied than the traditional software engineering approach explained above, namely one stressing prototyping, evolutionary development, learning and user involvement (Schuler and Namioka 1993).

In the Internet approach the standardisation process unifies the development of formal standards and their establishment as de facto ones. There is currently an interesting and relevant discussion going on about whether Internet's approach has reached its limits (see Eidnes 1994, p. 52; Steinberg 1995, p. 144). This is due to the fact that not only the technology changes. As the number of users grow, the organisation of the standardisation work also changes (Kahn 1994).

Contingency approaches

Due to the vast range of different "things" being standardized, it is unlikely that there is one best approach for all cases. Rather one needs a contingency approach in the sense that one needs to identify different approaches and criteria for choosing among them (David 1985?). Such criteria are the general maturity of the technology to be standardized, its range of diffusion, the number of actors involved, etc. This also implies that the approach used may change dynamically over time as the conditions for it changes. The evolution of the standardization approaches followed as the Internet has evolved illustrates such a contingency approach.

Branscomb and Kahin (op. cit.) hold that Internet demonstrates the remarkable potential (although perhaps the outer limits) for standards development and implementation in concert with rapid technological change. However, their view is that interoperability becomes harder to achieve as the functionality of the NII expands and draws in more and more vendors. They also say that it remains uncertain how well the Internet-style standards processes scale beyond its current reach. The difficulties in defining and implementing a new version of IP supports this view. This issue will be analysed extensively throughout this book.

Branscomb and Kahin expect that the U.S. Congress (and governments in other countries) will proceed with progressive deregulation of the telecommunication industry. This means that the importance of standardization will grow, while the government's authority to dictate standards will weaken. Yet these standards must not only accommodate the market failures of a deregulated industry; they must support the much more complex networks of the NII.

According to Branscomb and Kahin, all face the paradox that standards are critical to market development but, once accepted by the market, standards may threaten innovation, inhibit change, and retard the development of future markets. They conclude that these risks require standards processes to be future oriented. They consider Internet practices breaking down the dichotomy between anticipatory versus reactive standards by promoting an iterative standards development with concurrent implementation.

State of the art in research into standardisation

In activities aiming at implementing the NII and Bangemann reports, standards are identified as the key elements (ref.). However, it is becoming increasingly more accepted that current standardization approaches will not deliver (ref.). “Current approaches” means here the telecommunication model (including ISO’s). This approach is experienced to be too all too slow and inflexible. Some believe the democratic decision process is the problem, and that it should be replaced by more managerial government models (ref.).

Although the success of the Internet approach implies that it should be more widely adopted, as pointed out by Branscomb and Kahin, this approach has its limitations as well. This rather miserable state of affairs is motivating a growing research activity into standardization and standardization approach. We will here briefly point to some of the major resent and ongoing activities - activities this boot also is intended to contribute to.

ta med sts-aktige greier

digital libraries

Harvard project

hawkins & mansell - policical institutional economy

diffusjons teoretikere

CHAPTER 5 Openness and flexibility

Introduction

We will in this chapter look more closely at the open character of information infrastructures underscored in chapter 3. We will discuss and illustrate how this character is materialized and spell out (some of) its implications. We will in particular look at how openness generates requirements for flexible infrastructures.

At the core of this lies a dilemma. On the one hand, the expanding II supporting a growing population of users and new services accumulate pressure to make changes, but, on the other hand, this has to balanced against the conservative influence of the huge, already installed base of elements of the II. There is simply no way to accomplish abrupt changes to the whole II requiring any kind of overall coordination (for instance, so-called flag-days) because it is "too large for any kind of controlled rollout to be successful" (Hinden 1996, p. 63).

Standardization and flexibility are opposites. However, both are required. This fact makes the tension, interdependencies and interaction between these two issues crucial.

Openness

Defining openness

Infrastructures, as explained in chapter 3, are open in the sense that there is no limits for number of user, stakeholders, vendors involved, nodes in the network and other technological components, application areas, network operators, etc. This defining characteristic does not necessarily imply the extreme position that absolutely everything is included in every II. 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 have no connections.

The enabling character of infrastructures means that their use ares should not be predetermined or specified ahead of design and implementation. Enabling implies that infrastructures should be used where and as needed, as needs changes and use opportunities are discovered. This aspect of infrastructures is well illustrated by the growth in numbers of fields where the Internet is used, from the initial small group of researches to virtually everybody.

Each use area has its categories and groups of users. If numbers of use areas is unlimited, so is the number of users, which further makes it impossible to define strict limits for the number of components included in the infrastructure and the number of vendors and other stakeholders involved in its design and operation.

In the discussion about open systems (XXX Carl Hewitt), the term open usually means that a system is embedded in an environment and that the system cannot be properly understood without being considered as a part of its environment. Our use of the term open also includes this one.

Infrastructures are embedded in environments, upon which they are dependent. Change in the environment might require change in the infrastructure. Infrastructures are parts of their environments in a way making it difficult to define a strict border between infrastructure and environment. The infrastructure is a part of its environment just as the environment is a part of the infrastructure.

The Internet, for instance, is running over various basic telecommunication infrastructures like ATM and telephone networks. Are these networks a part of the Internet or not? In the same way are health care infrastructures (for instance patient record and telemedicine infrastructures) integrated with medical instruments (MR, CT, X-ray, ultrasound, ECG, EEG or endoscope). Are they a part of the infrastructure?

When using a specific information infrastructure, say an EDI network for transmission of lab reports or prescriptions, it will in fact be used in combination with other infrastructures like the ordinary telephone. In case of lab reports the telephone is used in emergency cases. In case of prescriptions will pharmacies call the general practitioner when there is something that requires clarification. The EDI network is not working unless it is combined with others in this way. The overall lab report and prescription transmission infrastructure also includes the telephone infrastructure.

Unlimited use

As mention above, the Internet is a good example of the impossibility of defining the areas and ways of using an infrastructure. Email was initially designed to support the coordination and management of the first version of the Internet. Today it is used in virtually all kinds of activities. The enabling character of infrastructures makes defining its use areas impossible. This implies further that defining user requirements is impossible.

Unlimited number of stakeholders

Unlimited use necessarily implies unlimited users as well as other stakeholders.The stakeholders around the design and use of an information infrastructure include at least:

- different categories of end-users;
- a collection of managers from involved user organizations;
- numerous vendors for various components of an information infrastructure;
- regulatory regimes (often governmental);
- standardisation bodies with professional standardisation experts and bureaucrats;
- telecom operators and service providers;
- political leaders (including Clinton) and institutions (including the European Union) rhetorically promoting visions of desirable changes in work and leisure;

The present dog-fight over Internet provides ample evidence for the crossing interests of the many stakeholders (Monteiro 1998). The last years have witnessed a transformation of Internet from an esoteric toy nurtured by a small, US based boy’s club to an infrastructure supporting a rich spectrum of users and interests. The traditional Internet community was small and cosy, consisting of academics and a collection of designers from large software vendors in the US. Only in 1992 did this traditional community become a minority user group in Internet. Today, there are a number of new stakeholders on the scene. The picture is highly dynamic and changes every day. With the widespread use of the WorldWideWeb, Internet opened up to a new, large and heterogeneous user community of young and old, experienced and non-experienced and technical and non-technical. Important elements of the design moved outside the control of the IETF. Especially the establishment of the W3C consortium and efforts to promote electronic transactions are influential. Key actors here are large software vendors (including IBM, Sun, Microsoft, Apple and HP) and financial institutions (including Visa and MasterCard). It is interesting, but regrettably beyond the present scope, to study the strategies of these vendors towards open standards. Microsoft, for instance, has traditionally been extremely "closed" in the sense that the different Microsoft products were hardwired into each other thus keeping competitors out (as did IBM during the golden age of main frames). Within Internet technology in general and Web in particular, areas where Microsoft has been lagging considerable behind, they are perfectly happy to be daring and explorative (XXXKILDE Haakon Lie 1996). They promote continued experimentation rather than early adoption of an existing solution, a solution developed by a competitor.

Another slow starting, but presumably influential, group of stakeholders are the telecom operators. Coming from a different paradigm altogether (Abbate 1994), telecom operators are increasingly trying to align with Internet rather than work against it. In Europe, for instance, telecom operators are key actors in the area of Internet access providers (REF Knut HS, Jarle B)). The media, especially TV and newspapers, struggle to come to grips with their role and use of Internet. Until now, however, they have played a marginal role.

A more technologically biased, but historically important, sense that information infrastructures involve numerous stakeholders concerns interoperability and portability. An information infrastructure is to avoid a too tight connection with a specific vendor. It should be multi-platform, ideally allowing a rich selection of operating systems, hardware architectures, applications and programming languages. XXFINN PASSE SITAT Needless to say, this is more easily stated that practised as illustrated in chapter 10.

Unlimited size of an information infrastructure

It is not given from the outset how big, that is, how widespread, an information infrastructure will be. It is an open, empirical question. Note that despite the apparent similarities, the situation is distinct from that in the product world. It is certainly true that nobody can tell for sure exactly many copies of MS Word will be sold, that the “size” of MS Word is unknown beforehand. But this has little or no impact on the design (beyond, of course, the increased pressure in terms of appealing functionality and appearance). There is no intrinsic connection between different stand-alone copies of MS Word. This is fundamentally different for an information infrastructure. As all nodes need to function together — regardless of their number — the design need to cater for this. It must be designed to be open to allow for an indefinite number of nodes to hook up.

An information infrastructure has to be designed to support a large, but unknown, population of users. So if hardwired constraints to further expansion surfaces, changes have to be made. These changes also take place during rapid diffusion. The very diffusion is in itself an important reason for the need for change. The number of hosts connected to Internet grew from about 1000 to over 300.000 during 1985-1991 (Smarr and Catlett 1992). The Matrix Information and Directory Services estimated the number to about 10 million in July 1995 (McKinney 1995).

A network with just a few hosts has to be designed differently from one with millions. When the number of hosts is open, the design has to be so as well.

The wider environment - a more rapidly changing world

Dominating accounts of today’s business and organizational reality is full of concepts like “globalization,” “competitiveness,” “flexibility,” and change. The “reality” depicted by these concepts are not isolated to the business world, it is rather creeping into virtually all our social life. Such accounts are found in popular press as well as scientific literature.

We are witnessing a rapidly increasing number of theoretical, speculative or empirical accounts dealing with the background, contents and implications for a restructuring of private and public organisations. The sources of these accounts mirror the complex and many-faceted issues raised of economical (OECD 1992), social (Clegg 1990), political (Mulgan 1991) and technological (Malone and Rockart 1993; Malone, Yates and Benjamin 1991) nature. A comprehensive account of this is practically prohibitive; the only feasible strategy is to focus attention on a restricted set of issues.

New organisational forms are assumed important in order to achieve enhanced productivity, competitiveness, flexibility, etc. New organisational forms are usually of a network type positioned between markets and hierarchies. The discussions about new organisational forms borrow from a number of sources.

From economics, basically relying on transaction-cost considerations, there is a growing pressure to accommodate to the “information economy” (Ciborra 1992; OECD 1992). Transaction-cost considerations fail to do justice to the dynamically changing division of labour and functions which are two important aspects of new organisational forms (Ciborra 1992). Within business policy literature the arguments focus on issues of innovation processes as facilitated through strategic alliances and globalization which emerge pragmatically from concerns about maintaining competitiveness in a turbulent environment (Porter 1990; von Hippel 1988). In organisational theory, one emphasises the weaknesses of centralised, bureaucratic control in terms of responsiveness to new situations (Clegg 1990). Ciborra (1992) sees new organisational forms as rational, institutional arrangements to meet the increased need for organisational learning. Technological development within information and communication technology are identified by some scholars as the driving force for the restructuring of organisations (Malone, Yates and Benjamin 1991; Malone and Rockart 1993).

Even such a brief exposition of theoretical considerations should make it evident that the issue of new organisational forms is vast. When we turn to what exists of empirical evidence, the picture gets even more complicated. This is because the empirical material document a far less clear-cut picture as it contains numerous contradicting trends (Applegate 1994; Capello and Williams 1992; Orlikowski 1991; Whitaker 1992).

A basic underlying theme in all these accounts is the view that our world is becoming increasingly more open as we are becoming more integrated. The integration is primarily due to improved transport and communication technology. As we are more tightly interwoven in a larger environment, our life is more open to influence by others - which again implies increased instability and unpredictability. The organizational response to this is increased flexibility and communication to more effectively adapt to and interoperate with the environment. In this sense, information and communication technology is both a cause and an effect of this trend, generating and continuously faster spinning spiral.

Implications: the need for flexible infrastructures

Systems and infrastructures

Information infrastructures are, of course, in a sense information systems. As such they share lots of properties. There is a general and fairly well-known argument that the use (at least the requirements) of an information system evolves over time. Boehm (198xx), for instance, includes this in his spiral model for systems development when explaining that systems development is really like aiming at a moving target. Requirements are neither complete nor stable. They are only gradually uncovered and they are dynamic. For larger systems it is also the case that it is impossible to foresee all relevant issues and problems, they are discovered as we go along, and the technology must be changed accordingly (Parnas and Clement 1986).

Hence, it is hardly news that requirements about use evolve. Still, mainstream systems development is biased towards early and fairly stable specifications (Pressman 1992). There is an alternative, much less widely applied, approach stressing prototyping, evolutionary development, learning and user involvement (Schuler and Namioka 1993). Systems development is viewed, in principle, as a mutual learning process where designers learn about the context of use and end-users about the technical possibilities (REF noe blautfisk greier). The first versions of a system are poor in quality compared to later ones. They are improved as users get experience in using them and discover what is needed as well as how the technology may be adapted to improved ways of working. For users it is impossible to tell in advance what kind of technology that will suit their needs best. User influence is an illusion unless it is based on experience of use.

The rationale for the importance of enabling a mutual learning process associated with systems development is, at least in parts, an ideologically biased one (REF empiriske studier). A less dogmatic approach, and one more immediately relevant to the development of information infrastructures, would be to inquire empirically into actual changes that have been implemented in information infrastructures in response to evolving patterns of use. This may be illustrated by a few of the changes of some OSI and Internet standards during their lifetime up till now.

OSI

OSI protocols have in fact been quite stable after their formal approval. This stability may to a large extent be explained by the fact that most OSI protocols did not diffuse. The OSI standard for e-mail, however, was approved in 1984. Four years later a new version came. It differed so much that a number of its features were incompatible with the earlier version (Rose 1992).

Internet

Internet has so far proved remarkably flexible, adaptable and extendable. It has undergone a substantial transformation — constantly changing, elaborating or rejecting its constituting standards — during its history. To keep track of all the changes, approximately quarterly a special report is issued which gives all the latest updates (RFC 1995). These changes also take place during rapid diffusion. The very diffusion is an important reason for the need for change. The number of hosts connected to Internet grew from about 1000 to over 300.000 during 1985-1991 (Smarr and Catlett 1992). The Matrix Information and Directory Services estimated the number to about 10 million in July 1995 (McKinney 1995).

The need for an II to continue to change alongside its diffusion is recognised by the designers themselves as expressed in an internal document describing the organisation of the Internet standardisation process: “From its conception, the Internet has been, and is expected to remain, an evolving system whose participants regularly factor new requirements and technology into its design and implementation” (RFC 1994, p. 6).

The IETF has launched a series of working groups which, after 4-5 years, are still struggling with different aspects of these problems. Some are due to new requirements stemming from new services or applications. Examples are asynchronous transmission mode, video and audio transmission, mobile computers, high speed networks (ATM) and financial transactions (safe credit card purchases). Other problems, for instance, routing, addressing and net topology, are intrinsically linked to and fuelled by the diffusion itself of Internet (RFC 1995). As commercial actors have been involved, the “triple A-s” - authentication, authorization, and accounting - have appeared as important issues. Until recently, when the Internet has been a non-commercial network, these issues have been non/exiting. For commercial service providers and network operators, they are absolutely necessary. This commercial turn requires that new features are added to a wide range of existing protocols.

The above illustrations clearly indicate that there is nothing which suggests that the pace or need for flexibility to change Internet will cease, quite the contrary (Smarr and Catlett 1992; RFC 1994, 1995).

During the period between 1974 and 1978 four versions of the bottom-most layer of the Internet, that is, the IP protocol were developed and tested out (Kahn 1994). For almost 15 years it has been practically stable. It forms in many respects the core of the Internet by providing the basic services which all others build upon (cf. our earlier description). An anticipated revision of IP is today the subject of “spirited discussions” (RFC 1995, p. 5). The discussions are heated because the stakes are high. The problems with the present version of IP are acknowledged to be so grave that Internet, in its present form, cannot evolve for more than an estimated 10 years without ceasing to be a globally, inter-connected network (ibid., pp. 6-7; Eidnes 1994, p. 46). This situation is quite distinct from the more popular conception of an inevitable continued development of Internet. There are a whole set of serious and still unresolved problems. Among the more pressing ones, there is the problem that the “address space” will run out in few years. The Internet is based on the fact that all nodes (computers, terminals and printers) are uniquely identified by an address. This size of this space is finite and determined by how one represents and assigns addresses. The problem with exhausting the current address space is serious as it will block any further diffusion of Internet for the simple reason that there will not be any free addresses to assign to new nodes wishing to hook up. The difficulty is that if one switches to a completely different way of addressing, one cannot communicate with the “old” Internet. One is accordingly forced to find solutions which allow both the “old” (that is, the present) version of IP to function alongside the new and non-existing IP.

This case is elaborated at greater length in chapter 10 (with a full account in (Monteiro 1998)). A core dilemma is how to balance the urge for making changes against the conservative influence of the installed base (see chapter 9). As a standard is implemented and put into widespread use, the effort of changing it increases accordingly simply because any changes need to be propagated to a growing population of geographically and organisationally dispersed users as captured by the notion of “network externalities” (Antonelli 1993, Callon 1994, p. 408) or the creation of lock-ins and self-reinforcing effects (Cowan 1992, pp. 282-283).

As the components of IIs are inter-connected, standardisation sometimes requires flexibility in the sense that to keep one component standardised and stable others must change. Enabling mobile computers network connections, for instance, requires new features to be added to IIs (Teraoka et al. 1994). These may be implemented either as extensions to the protocols at the network, transport or application level of the OSI model. If one wants to keep one layer stable others must change.

Enabling flexibility to change

The primary principle for enabling flexibility is modularization, i.e. "black-boxing." Modularization as a strategy for coping with design is employed by most engineers, not only those involved with II (Hård 1994). It could, however, be maintained that in the case of computer science (including the development of II) modularization is systematically supported through a large and expanding body of tools, computer language constructs and design methodologies. Elaborating this would carry us well beyond the scope of this paper, but it is indeed possible to present the historical development of a core element of computer science, namely the evolution of programming language, as very much influenced with exactly how to find constructs which support flexibility to change in the long run by pragmatically deciding how to restrict or discipline local flexibility. The interested reader might want to recast, say, the controversy over structured programming along these lines, that is, recognising the call for structured constructs as a means for enabling flexibility in the long run by sacrificing local flexibility of the kind the GOTO statement offers. (The GOTO statement offers great flexibility in how to link micro level modules together at the cost of diminishing the flexibility to change the modules later on.)

Decomposition and modularization are at the same time a basis for flexibility in II: flexibility presupposes modularization. The reason for this, at least on a conceptual level, is quite simple. The effect of black-boxing is that only the interface (the outside) of the box matters. The inside does not matter and may accordingly be changed without disturbing the full system provided the interface looks the same. As long as a box is black, it is stable and hence standardised. In this sense *standardisation is a precondition for flexibility.*

Two forms of this modularization need to be distinguished. Firstly, it may give rise to a layered or hierarchical system. OSI's seven layered communication model provides a splendid example of this. Each layer is uniquely determined through its three interfaces: the services it offers to the layer immediately above, the services it uses in the layer immediately below and the services a pair of sender and receiver on the same level make use of.

Secondly, modularization may avoid coupling or overlap between modules by keeping them "lean". One way this modularization principle is applied is by defining mechanisms for adding new features without changing the existing ones. In the new version of IP, for instance, a new mechanism is introduced to make it easier to define new options (RFC 1995). Another example is the WorldWideWeb which is currently both diffusing and changing very fast. This is possible, among other reasons, because it is based on a format defined such that one implementation simply may skip or read as plain text elements it does not understand. In this way, new features can be added so old and new implementations can work together.

Hampering flexibility to change

There are three principal ways the flexibility to change an II is hampered. Breaking either of the two forms of modularization enabling flexibility described above accounts for two of the three ways flexibility is hampered. To illustrate how lack of hierarchical modularization may hamper flexibility, consider the following instance of a violation found in OSI. In the application level standard for e-mail, the task of uniquely identifying a person is not kept apart from the conceptually different task of implementing the way a person is located. This hampers the flexibility because if an organisation changes the way its e-mail system locates a person (for instance, by changing its network provider), all the unique identifications of the persons belonging to the organisation have to be changed as well. Most OSI protocols are good illustrations of violations of the "lean-ness" principle. Although the OSI model is an excellent example of hierarchical modularization, each OSI protocol is so packed with features that they are hardly possible to implement and even harder to change (Rose 1992). The reason is simply that it is easier to change a small and simple component than a large and complex one. Internet protocols are much simpler, that is, leaner, that OSI ones, and accordingly easier to change.

The third source of hampered flexibility is the diffusion of the II. As a standard is implemented and put into widespread use, the effort of changing it increases accordingly simply because any changes need to be propagated to a growing population of geographically and organisationally dispersed users as captured by the notion of "network externalities" (Antonelli 1993, Callon 1994, p. 408) or the creation of lock-ins and self-reinforcing effects (Cowan 1992, pp. 282-283).

At the moment, Internet appears to be approaching a state of irreversibility. Consider the development of a new version of IP described earlier. One reason for the difficulties in developing a new version of IP is the size of the installed base of IP protocols which must be replaced while the network is running (cf. rate of diffusion cited earlier). Another major difficulty stems from the inter-connectivity of standards: a large number of other technical components depend on IP. An internal report assesses the situation more precisely as: "Many current IETF standards are affected by [the next version of] IP. At least 27 of the 51 full Internet Standards must be revised (...) along with at least 6 of the 20 Draft Standards and at least 25 of the 130 Proposed Standards." (RFC 1995, p. 38).

The irreversibility of II has not only a technical basis. An II turn irreversible as it grows due to numbers of and relations between the actors, organisations and institutions involved. In the case of Internet, this is perhaps most evident in relation to new, commercial services promoted by organisations with different interests and background. The transition to the new version of IP will require coordinated actions from all of these parties. It is a risk that “everybody” will await “the others” making it hard to be an early adopter. As the number of users as well as the types of users grow, reaching agreement on changes becomes more difficult (Steinberg 1995).

The interdependencies between standardization and flexibility

We have sketched, drawing on both conceptual arguments and empirical illustrations, in what sense an information infrastructure needs to be open-ended and flexible. However, the major difficulty may be to replace one working version with another working one as change will introduce some kind of incompatibility which may cause a lock-in situation. As the components of information infrastructures are inter-connected, *standardisation requires flexibility* in the sense that to keep one component standardised and stable others must change. Enabling mobile computers network connections, for instance, requires new features to be added to IIs (Teraoka et al. 1994). These may be implemented either as extensions to the protocols at the network, transport or application level of the OSI model. If one wants to keep one layer stable others must change. This aspect of information infrastructure we might call anticipated and alternating flexibility.

There are, as we have seen in earlier chapters, lots of Internet standards. These standards do not all fit into a tidy, monolithic form. Their inter-relationships are highly complex. Some are organised in a hierarchical fashion as illustrated by the bulk of standards from OSI and Internet as outlined below. Others are partly overlapping as, for instance, in the case where application specific or regional standards share some but not all features. Yet others are replaced, wholly or only in part, by newer standards creating a *genealogy of standards*. This implies that the interdependencies of the totality of standards related to II form a complex network. The heterogeneity of II standards, the fact that one standard includes, encompasses or is intertwined with a number of others, is an important aspect of II. It has, we argue, serious implications for how the tension between standardisation and flexibility unfolds in II.

We have in this chapter focused on just one type of flexibility — flexibility to change. Another type of flexibility is flexibility in use. This means that the information infrastructure may be used in many different ways, serving different purposes as it is — without being changed. This is exactly what gives infrastructures their enabling function. Stressing this function makes use flexibility crucial. Flexibility of use and change are linked in the sense that increased flexibility of use decreases the need for flexibility for change and vice versa. An aspect of both flexibility of use and change is to provide possibilities for adaptation to different local needs and practices, avoiding unacceptable constraints being imposed by some centralized authority. Flexibility of use is the topic of chapter 7 when we develop the notion of an "inscription" to talk about the way material (or non-material) artefacts (attempt to) shape behaviour. To do so, however, it is necessary to frame this within a broader understanding of the interplay between artefacts and human behaviour. To this end, we in the following chapter 6 develop a theoretical framework called actor-network theory (ANT).

CHAPTER 6

Socio-technical webs and actor network theory

Introduction

We now turn to the socio-technical nature of information infrastructures. As outlined in chapters 3 and 4 above, the development of an information infrastructure needs to be recognised as an ongoing socio-technical negotiation. An analysis of it accordingly presumes a suitable vehicle. This chapter contains a presentation of one such vehicle, namely actor-network theory. It is intended to pave the road for describing and analysing how issues of flexibility and standardisation outlined in chapters 4 and 5 unfold in relation to information infrastructure. This chapter accordingly functions as a stepping stone for the chapters than follow.

Technology and society: a brief outline

The relationship between technology and society may be conceptualised in many ways. We embrace the fairly widespread belief that IT is a, perhaps the, crucial factor as it simultaneously enables and amplifies the currently dominating trends for restructuring of organisations (Applegate 1994; Orlikowski 1991). The problem, however, is that this belief does not carry us very far; it is close to becoming a cliche. To be instructive in an inquiry concerning current organisational transformations, one has to supplement it with a grasp of the interplay between IT and organisations in more detail. We need to know more about how IT shapes, enables and constrains organisational changes. Two extreme end points of a continuum of alternatives are, on the one hand, technological determinism holding that the development of technology follows its own logic and that the technology determine its use (Winner 1977) and, on the other hand, social reductionism or constructionism (Woolgar 1991), (which comes close to technological somnambulism (Pfaffenberger 1988; Winner 1977)) holding that society and its actors develop the technology it “wants” and use it as they want, implying that technology in itself plays no role. A series of Braverman inspired studies appeared in the late 70s and early 80s biased towards a technological determinist position arguing that the use of IT was but the latest way of promoting management’s interests regarding deskilling and control of labour (Sandberg 1979). Later, a number of studies belonging close to the social constructivist end of the continuum were produced which focused on diversity of use among a group of users and displaying use far beyond what was anticipated by the designers (Henderson and Kyng 1991; Woolgar 1991b).

A more satisfactory account of the interwoven relationship between IT and organisational transformations is lacking. More specifically, we argue that we need to learn more about how this interplay works, not only that it exists. This implies that it is vital to be more concrete with respect to the specifics of the technology. As an information system (IS) consists of a large number of modules and inter-connections, it may be approached with a varying degree of granularity. We cannot indiscriminatingly refer to it as IS, IT or computer systems. Kling (1991, p. 356) characterises this lack of precision as a “convenient fiction” which “deletes nuances of technical differences”. It is accordingly less than prudent to discuss IS at the granularity of an artefact (Pfaffenberger 1988), the programming language (Orlikowski 1992), the overall architecture (Applegate 1994) or a media for communication (Feldman 1987). To advance our understanding of the interplay it would be quite instructive to be as concrete about which aspects, modules or functions of an IS enable or constrain which organisational changes — without collapsing this into a deterministic account (Monteiro, Hanseth and Hatling 1994).

Today, the majority of scholars in the field adhere to an intermediate position somewhere between the two extreme positions outlined above. The majority of accounts end up with the very important, but all too crude, insight that “information technology has both restricting and enabling implications” (Orlikowski and Robey 1991, p. 154). This insight — that IT enables and constrains — is reached using a rich variety of theoretical frameworks including structuration theory (Orlikowski and Robey 1991), phenomenology (Boland and Greenberg 1992), hermeneutics (Klein and Lyytinen 1992) or Habermas’ theory of communicative action (Gustavsen and Engelstad 1990).

Hence, there can hardly be said to be a lack of suggestions for suitable theoretical frameworks (Kling 1991; Monteiro and Hanseth 1995). We will, however, introduce yet another one, actor network theory, which we believe will bring us one step further towards a more detailed understanding of the relationships between information technology and its use (Akrich 1992; Akrich and Latour 1992; Callon 1991, 1994; Latour 1987). This choice is motivated by the way actor network theory, especially in the minimalistic variant we employ, offers a language for describing the many small, concrete technical and non-technical mechanisms which go into the building and use of information infrastructures. (We will particularly look at the negotiations of standards.) Actor network theory accordingly goes a long way in describing which and how actions are enabled and constrained.

In this chapter we develop a minimalistic vocabulary of ANT intended to be a manageable, working way of talking about use and design of IT (see (Walsham 1997 ifip8.2) for a survey of the use of ANT in our field). As a frame of reference, we first sketch and compare alternatives to ANT. We do not want to be seen as too dogmatic. There certainly exist fruitful alternatives and placing ANT in a wider context of related thinking is relevant.

The duality of technology

The problem of how to conceptualise and account for the relationship between, on the one hand, IT development and use and, on the other hand, organisational changes is complex — to say the least. A principal reason for the difficulty is due to the contingent, interwoven and dynamic nature of the relationship. There exists a truly overwhelming body of literature devoted to this problem. We will discuss a selection of contributions which are fairly widely cited and we which consider important. (Consult, for instance, (Coombs, Knights and Willmott 1992; Kling 1991; Orlikowski and Robey 1991; Walsham 1993) for a broader discussion.) Our purpose is to motivate a need to incorporate into such accounts a more thorough description and understanding of the minute, grey and seemingly technical properties of the technology and how these are translated into non-technical ones.

The selection of contributions we consider all acknowledge the need to incorporate, in one way or another, that subjects interpret, appropriate and establish a social construction of reality (Galliers 1992; Kling 1991; Orlikowski 1991; Orlikowski and Robey 1991; Smithson, Baskerville and Ngwenyama 1994; Walsham 1993). This alone enables us to avoid simple-minded, deterministic accounts. The potential problem with a subjectivist stance is how to avoid the possibility that, say, an IS could be interpreted and appropriated completely freely, that one interpretation would be just as reasonable as any other. This position obviously neglects the constraining effects the IS have on the social process of interpretation (Akrich 1992; Bijker 1993; Orlikowski and Robey 1991). In other words, it is absolutely necessary to recognise the "enabling and constraining" abilities of IT. A particularly skilful and appealing elaboration of this insight is the work done by Orlikowski, Walsham and others building on Giddens' structuration theory (Orlikowski 1991, 1992; Orlikowski and Robey 1991; Walsham 1993).

Despite the fact that these accounts, in our view, are among the most convincing conceptualisations, they have certain weaknesses. These weaknesses have implications for the way we later on will approach the question of the relationship between information infrastructure and new organisational forms. Our principal objection to conceptualisations like (Orlikowski 1991, 1992; Orlikowski and Robey 1991; Walsham 1993) is that they are not fine-grained enough with respect to the technology to form an appropriate basis for understanding or to really inform design. Before substantiating this claim, it should be noted that the studies do underline an important point, namely that "information technology has both restricting and enabling implications" (Orlikowski and Robey 1991, p. 154). We acknowledge this, but are convinced that it is necessary to push further: to describe in some detail how and where IT restricts and enables action. At the same time, we prepare the ground for the alternative framework of ANT describing the social construction of technology. To this end, we briefly sketch the position using structuration theory.

The aim of structuration theory is to account for the interplay between human action and social structures. The notion of "structure" is to be conceived of as an abstract notion; it need not have a material basis. The two key elements of structuration theory according to Walsham (1993, p. 68) are: (i) the manner in which the two levels of actions and structure are captured through the duality of structure, and (ii) the identification of modalities as the vehicle which link the two levels. One speaks of the duality of structure because structure constrains actions but, at the same time structures are produced (or more precisely: reproduced and transformed) through human action. This mutual interplay is mediated through a linking device called modalities. As modalities are what link action and structure, and their relationship is mutual, it follows that these modalities operate both ways.

There are three modalities: *interpretative scheme, facility* and *norm*. An interpretative scheme deals with how agents understand and how this understanding is exhibited. It denotes the shared stock of knowledge which humans draw upon when interpreting situations; it enables shared meaning and hence communication. It may also be the reason why communication processes are inhibited. In applying this framework to IT, Orlikowski and Robey (1991, p. 155) notes that "software technology conditions certain social practices, and through its use the meanings embodied in the technology are themselves reinforced". The second modality, facility, refers to the mobilisation of resources of domination, that is, it comprises the media through which power is exercised. IT, more specifically, "constitutes a system of domination" (ibid., p. 155). The third modality, norms, guide action through mobilisation of sanctions. As a result, they define the legitimacy of interaction. They are created through continuous use of sanctions. The way this works for IT is that IT "codifies" and "conveys" norms (ibid., pp. 155 - 156).

Given this admittedly brief outline of the use of structuration theory for grasping IT, we will proceed by documenting in some detail how these accounts fail to pay proper attention to the specifics of IT. Orlikowski and Robey (1991, p. 160) point out how "tools, languages, and methodologies" constrain the design process. The question is whether this lumps too much together, whether this is a satisfactory level of precision with regards to the specifics of IT. There is, after all, quite a number of empirical studies which document how, say, a methodology fails to constrain design practice to any extent; it is almost never followed (Curtis, Krasner and Iscoe 1988, Ciborra and Lanzara 1994). Referring to Orlikowski (1991), Walsham (1993, p. 67) notes that "the ways in which action and structure were linked are only briefly outlined". This is hardly an overstatement as (Orlikowski 1991), as opposed to what one might expect from examining "in detail the world of systems development" (ibid., p. 10), maintains that the CASE tool — which is never described despite the fact that such tools exhibit a substantial degree of diversity (Vessey, Jarvenpaa and Tractinsky 1992) — was the "most visible manifestation" of a strategy to "streamline" the process (Orlikowski 1991, p. 14). (Orlikowski 1992) suffers from exactly the same problem: organisational issues are discussed based on the introduction of Lotus Notes. We are never explained in any detail, beyond referring to it as "the technology" or "Notes", the functions of the applications. This is particularly upsetting considering the fact that Lotus Notes is a versatile, flexible application level programming language.

In instructive, in-depth case studies, Walsham (1993) does indeed follow up his criticism cited above by describing in more detail than Orlikowski (1991, 1992) how the modalities operate. But this increased level of precision does not apply to the specifics of the technology. The typical level of granularity is to discuss the issue of IBM vs. non-IBM hardware (ibid., pp. 92-94), centralised vs. decentralised systems architecture (ibid., p. 105) or top-down, hierarchical control vs. user-control (ibid., p. 136 - 138).

Not distinguishing more closely between different parts and variants of the elements of the IS is an instance of the aforementioned "convenient fiction" (Kling 1991, p. 356). An unintended consequence of not being fine-grained enough is removing social responsibility from the designers (ibid., p. 343). It removes social responsibility in the sense that a given designer in a given organisation obliged to use, say, a CASE tool, may hold that it is irrelevant how she uses the tool, it is still a tool embodying a certain rationale beyond her control.

What is required, as already mentioned, is a more detailed and fine-grained analysis of the many mechanisms, some technical and some not, which are employed in shaping social action. We are not claiming that structuration theory cannot deliver this (cf. Walsham 1993, p. 67). But we are suggesting that most studies conducted so far (Korpela 1994; Orlikowski 1991, 1992; Orlikowski and Robey 1991; Walsham 1993) are lacking a description, at a satisfactory level of precision, of how specific elements and functions of an IS relate to organisational issues. We also suggest that the framework provided by actor-network theory (ANT) is more promising in this regard. We proceed to give an outline of the basics of ANT based on (Akrich 1992; Akrich and Latour 1992; Callon 1991; Latour 1987) before discussing what distinguishes it from the position outlined above.

Actor network theory - a minimalistic vocabulary

ANT was born out of ongoing efforts within the field called social studies of science and technology. It was not intended to conceptualise information technology as such, and certainly not the ongoing design of a technology like information infrastructure, that is a technology in the process of being developed.

The field of social studies of technology in general and ANT in particular are evolving rapidly (REFS). It is a task in itself keeping up with the latest developments. Our aim is modest: to extract a small, manageable and useful vocabulary suited an adequate understanding of the challenges of developing information infrastructure. To this end, we simplify and concentrate on only the aspects of ANT most relevant to our endeavour.

What is an actor network, anyway?

The term "actor network", the A and N in ANT, is not very illuminating. It is hardly obvious what the term implies. The idea, however, is fairly simple. When going about doing your business — driving your car or writing a document using a word-processor — there are a lot of things that influence how you do it. For instance, when driving a car, you are influenced by traffic regulations, prior driving experience and the car's manoeuvring abilities, the use of a word-processor is influenced by earlier experience using it, the functionality of the word-processor and so forth. All of these factors are related or connected to how you act. You do not go about doing your business in a total vacuum but rather under the influence of a wide range of surrounding factors. The act you are carrying out and all of these influencing factors should be considered together. This is exactly what the term actor network accomplishes. An actor network, then, is the act linked together with all of its influencing factors (which again are linked), producing a network.[1]

An actor network consists of and links together both technical and non-technical elements. Not only the car's motor capacity, but also your driving training, influence your driving. Hence, ANT talks about the heterogeneous nature of actor networks. In line with its semiotic origin, actor network theory is granting all entities of such a heterogeneous network the same explanatory status as "semiotics is the study of order building (...) and may be applied to settings, machines, bodies, and programming languages as well as text (...) [because] semiotics is not limited to signs" (Akrich and Latour 1992, p.259). It might perhaps seem a radical move to grant artefacts the same explanatory status as human actors: does not this reduce human actors to mere objects and social science to natural science? We intend to bracket this rather dogmatic issue. Interested readers should consult (Callon and Latour 1992; Collins and Yearley 1992). For our purposes, what is important is that this move has the potential for increasing the level of detail and precision. More specifically, allowing oneself not to distinguish a priori between social and technical elements of a socio-technical web encourages a detailed description of the concrete mechanisms at work which glue the network together — without being distracted by the means, technical or non-technical, of actually achieving this. If really interested in discovering influential factors regarding the way you drive, we should focus on what turns out to be actually influential, be it technical (the motor' capacity) or non-technical (your training).

Inscription and translation

Two concepts from actor network theory are of particular relevance for our inquiry: *inscription* (Akrich 1992; Akrich and Latour 1992) and *translation* (Callon 1991, 1994; Latour 1987). The notion of inscription refers to the way technical artefacts embody patterns of use: "Technical objects thus simultaneously embody and measure a set of relations between heterogeneous elements" (Akrich 1992, p. 205). The term inscription might sound somewhat deterministic by suggesting that action is inscribed, grafted or hard-wired into an artefact. This, however, is a misinterpretation. Balancing the tight-rope between, on the one hand, an objectivistic stance

1. One way of putting this is that the actor network spells out the contents of the context or situatedness of an action (Suchman 1987).

where artefacts determine the use and, on the other hand, a subjectivistic stance holding that an artefact is always interpreted and appropriated flexibly, the notion of an inscription may be used to describe how concrete anticipations and restrictions of future patterns of use are involved in the development and use of a technology. Akrich (1992, p. 208, emphasis added) explains the notion of inscription in the following way:

> Designers thus define actors with specific tastes, competencies, motives, aspirations, political prejudices, and the rest, and they assume that morality, technology, science, and economy will evolve in particular ways. *A large part of the work of innovators is that of "inscribing" this vision of (or prediction about) the world in the technical content of the new object*. (...) The technical realization of the innovator's beliefs about the relationship between an object and its surrounding actors is thus an attempt to predetermine the settings that users are asked to imagine (...).

Stability and social order, according to actor network theory, are continually negotiated as a social process of aligning interests. As actors from the outset have a diverse set of interests, stability rests crucially on the ability to *translate*, that is, re-interpret, re-present or appropriate, others' interests to one's own. In other words, with a translation one and the same interest or anticipation may be presented in different ways thereby mobilising broader support. A translation presupposes a medium or a "material into which it is inscribed", that is, translations are "embodied in texts, machines, bodily skills [which] become their support, their more or less faithful executive" (Callon 1991, p. 143).

In ANT terms, design is translation - "users'" and others' interests may, according to typical ideal models, be translated into specific "needs," the specific needs are further translated into more general and unified needs so that these needs might translated into one and the same solution. When the solution (system) is running, it will be adopted by the users by translating the system into the context of their specific work tasks and situations.

In such a translation, or design, process, the designer works out a scenario for how the system will be used. This scenario is inscribed into the system. The inscription includes *programs of action* for the users, and it defines roles to be played by users and the system. In doing this she is also making implicit or explicit assumptions about what competencies are required by the users as well as the system. In ANT terminology, she delegates roles and competencies to the components of the socio-technical network, including users as well as the components of the system (Latour

1991). By inscribing programs of actions into a piece of technology, the technology becomes an actor[2] imposing its inscribed program of action on its users.

The inscribed patterns of use may not succeed because the actual use deviates from it. Rather than following its assigned program of action, a user may use the system in an unanticipated way, she may follow an anti-program (Latour 1991). When studying the use of technical artefacts one necessarily shifts back and forth "between the designer's projected user and the real user" in order to describe this dynamic negotiation process of design (Akrich 1992, p. 209).

Some technologies inscribe weak/flexible programs of action while others inscribe strong/inflexible programs. Examples of the former are tools, the hammer being a classic example, and the assembly line of Chaplin's "Modern times" a standard illustration of the latter.

Inscriptions are given a concrete content because they represent interests inscribed into a material. The flexibility of inscriptions vary, some structure the pattern of use strongly, others weakly. The strength of inscriptions, whether they must be followed or can be avoided, depends on the *irreversibility* of the actor-network they are inscribed into. It is never possible to know before hand, but by studying the sequence of attempted inscriptions we learn more about exactly how and which inscriptions were needed to achieve a given aim. To exemplify, consider what it takes to establish a specific work routine. One could, for instance, try to inscribe the routine into required skills through training. Or, if this inscription was too weak, one could inscribe the routine into a textual description in the form of manuals. Or, if this still is too weak, one could inscribe the work routines by supporting them by an information system. Hence, through a process of translation, one and the same work routine may be attempted inscribed into components of different materials, components being linked together into a socio-technical network. By adding and superimposing these inscriptions they accumulate strength.

Latour (1991) provides an illuminating illustration of this aspect of actor network theory. It is an example intended for pedagogic purposes. Hotels, from the point of view of management, want to ensure that the guests leave their keys at the front desk when leaving. The way this objective may be accomplished, according to actor network theory, is to inscribe the desired pattern of behaviour into an actor-network. The question then becomes how to inscribe it and into what. This is impossible to know for sure before hand, so management had to make a sequence of trials to test the strength of different inscriptions. In Latour's story, management

2. Or "actant" as would be the more precise term in actor network theory (Akrich and Latour 1992).

first tried to inscribe it into an artifact in the form of a sign behind the counter requesting all guests to return the key when leaving. This inscription, however, was not strong enough. Then they tried having a manual door-keeper — with the same result. Management then inscribed it into a key with a metal knob of some weight. By stepwise increasing the weight of the knob, the desired behaviour was finally achieved. Hence, through a succession of translations, the hotels' interest were finally inscribed into a network strong enough to impose the desired behaviour on the guests.

Four key aspects of inscriptions

There are four aspects of the notions of inscription and translation which are particularly relevant and which we emphasise in our study: (i) the identification of explicit *anticipations* (or scenarios) of use held by the various actors during design (that is, standardisation), (ii) *how* these anticipations are translated and inscribed into the standards (that is, the materials of the inscriptions), (iii) *who* inscribes them and (iv) the *strength* of these inscriptions, that is, the effort it takes to oppose or work around them.

Irreversibility

A key feature of information infrastructure, outlined in chapter 5, is the difficulty of making changes. Using and extending the core ANT vocabulary developed above, this vital aspect may be lifted forward to occupy centre stage. In ways to be elaborated in greater detail in subsequent chapters, an in information infrastructure is an aligned actor network. The constitutive elements of an information infrastructure — the collection of standards and protocols, user expectations and experience, bureaucratic procedures for passing standards — inscribe patterns of use. But is it not possible to express this more precisely, to somehow "measure" the net effects (a dangerous expression, but let it pass) to which these superimposed inscriptions actually succeed in shaping the pattern of use, to "measure" the strength of an inscription?

Callon's concept of the (possible) irreversibility of an aligned network captures the accumulated resistance against change quite nicely (Callon 1991, 1992, 1994). It describes how translations between actor-networks are made durable, how they can resist assaults from competing translations. Callon (1991, p. 159) states that the degree of irreversibility depends on (i) the extent to which it is subsequently impossible to go back to a point where that translation was only one amongst others and (ii) the extent to which it shapes and determines subsequent translations.

The notions which at the present stage in our analysis pay most adequate justice to the accumulating resistance against change, and the tight inter-connection between different parts of an II are alignment, irreversibility and accordingly momentum (Hughes and Callon both underline the similarities with the other, see Callon 1987, p. 101; Hughes 1994, p. 102). Despite their ability to account for the anticipated and interleaved flexibility of an II, these notions down-play this phenomenon to the point of disappearance. To make this point more precise, consider the notion of momentum which Hughes (1994) discusses as a possible candidate for conceptualising the development of infrastructure technologies.

The crucial difference between Hughes and Callon is connected with how the dynamics of momentum unfolds. Hughes describes momentum as very much a self-reinforcing process gaining force as the technical system grows "larger and more complex" (ibid., p. 108). It is reasonable to take the rate of diffusion of Internet during recent years as an indication of its considerable momentum. Major changes which seriously interfere with the momentum are, according to Hughes, only conceivable in extraordinary instances: "Only a historic event of large proportions could deflect or break the momentum [of the example he refers to], the Great Depression being a case in point" (ibid., p. 108) or, in a different example, the "oil crises" (ibid., p. 112). This, however, is not the case with II. As illustrated with the issue of the next version of IP in Internet, radical changes are regularly required and are to a certain extent anticipated.Momentum and irreversibility are accordingly contradictory aspects of II in the sense that if momentum results in actual — not only potential — irreversibility, then changes are impossible and it will collapse. Whether the proposed changes in Internet are adequate and manageable remains to be seen.

Actor networks meet structuration theory

Having given an outline of ANT, let us turn to see what is achieved *vis-à-vis* structuration theory. The principal improvement, as we see it, is the ability ANT provides to be more specific and concrete with respects to the functions of an IS. It is not the case, in our view, that ANT in every respect is an improvement over structuration theory. We only argue that it applies to the issue of being specific about the technology. For instance, we consider the important issue of the structuring abilities of institutions to be better framed within structuration theory than within ANT. Let us explain why we consider it so. We first compare the two theories on a general level, partly relying on pedagogic examples. Then we attempt to reinterpret (Orlikowski 1991) in terms of ANT.

Inscriptions are given a concrete content because they represent interests inscribed into a material. The flexibility of inscriptions vary, that is, some structure the pattern of use strongly, others quite weakly. The power of inscriptions, that is, whether they must be followed or can be avoided, depends on the irreversibility of the actor-network they are inscribed into. It is never possible to know before hand, but by studying the sequence of inscriptions we learn more about exactly how and which inscriptions were needed to achieve a given aim. To exemplify, consider what it takes to establish a specific work routine. One could, for instance, try to inscribe the required skills through training. Or, if this inscription was too weak, one could inscribe into a textual description of the routines in the form of manuals. Or, if this still is too weak, one could inscribe the work routines by supporting them by an IS.

ANT's systematic blurring of the distinction between the technical and the non-technical extends beyond the duality of Orlikowski and Robey (1991) and Walsham (1993). The whole idea is to treat situations as essentially equal regardless of the means; the objective is still the same. Within ANT, technology receives exactly the same (explanatory!) status as human actors; the distinction between human and non-human actors is systematically removed. ANT takes the fact that, in a number of situations, technical artefacts in practice play the same role as human actors very seriously: the glue which keeps a social order in place is a heterogeneous network of human and non-human actors. A theoretical framework which makes an a priori distinction between the two is less likely to manage to keep its focus on the aim of a social arrangement regardless of whether the means for achieving this are technical or non-technical. The consequence of this is that ANT supports an inquiry which traces the social process of negotiating, redefining and appropriating interests back and forth between an articulate explicit form and a form where they are inscribed within a technical artefact. With reference to the small example above, the inscriptions attempting to establish the work routine were inscribed in both technical and non-technical materials. They provide a collection of inscriptions — all aimed at achieving the same effect — with a varying power. In any given situation, one would stack the necessary number of inscriptions which together seem to do the job.

We believe that the empirical material presented by Orlikowski (1991) may, at least partially, be reinterpreted in light of ANT.[3] Her primary example is the development and use of a CASE tool in an organisation she calls SCC. The control (and productivity) interests of management are inscribed into the tool. The inscriptions

3. Doing this in any detail would, of course, demand access to the empirical material beyond the form in which it is presented in the article. We have no such access.

are so strong that the consultants do as intended down to a rather detailed level. The only exceptions reported are some senior consults saying that they in some rare instances do not do as the tool require.

What is missing, then, in comparison with ANT is to portray this as more a step-wise alignment than the kind of all-in-one character of (ibid.). In ANT terms, the management's control interests are inscribed into the CASE tool in forms of detailed inscriptions of the consultants behaviour. The inscriptions are very strong in the sense that there is hardly any room for interpretive flexibility. The CASE tool is the result of a long process where management's control and productivity interests have been translated into a larger heterogeneous actor-network encompassing career paths, work guidelines, methodologies and, finally, the CASE tool. Together these elements form an actor-network into which consultants' behaviour are inscribed. Just like Latour's example presented above, the inscriptions become stronger as they are inscribed into a larger network. This network is developed through successive steps where inscriptions are tested out and improved until the desired outcome is reached. It is only when, as a result of a long sequence of testing and superpositioning of inscriptions, that one ends up in situations like the one presented by Orlikowski (1991). If one succeeds in aligning the whole actor-network, the desired behaviour is established. Analytically, it follows from this that if any one (or a few) of the elements of such an actor-network is not aligned, then the behaviour will not be as presented by Orlikowski (ibid.). Empirically, we know that more often than not the result is different from that of Orlikowski's case (Curtis, Krasner and Iscoe 1988; Vessey, Jarvenpaa and Tractinsky 1992).

We end this section by merely pointing out another issue we find problematic with (Orlikowski 1992). She states that "[t]he greater the spatial and temporal distance between the construction of a technology and its application, the greater the likelihood that the technology will be interpreted and used with little flexibility. Where technology developers consult with or involve future users in the construction and trial stages of a technology, there is an increased likelihood that it will be interpreted and used more flexibly" (ibid., p. 421). We agree on the importance of user participation in design. According to ANT, however, the interpretive flexibility of a technology may increase as the distance between designers and users increases. Interpretive flexibility means unintended use, i.e. using the technology different for what is inscribed into it. When the designers are close to the users, the network into which the intended user behaviour is inscribed will be stronger and accordingly harder for the users not to follow this. An important aspect of ANT is its potential to account for how restricted interpretative flexibility across great distances can be obtained (Law 1986).

The chapters which follow build upon this minimalistic vocabulary of ANT. The notion of an actor-network will implicitly be assumed throughout. The notion of an inscription is closely linked to the (potential lack of) flexibility of an information infrastructure and will be elaborated in chapter 7. It also plays a central role when describing and critically assessing the prevailing approaches to the design of information infrastructures in chapter 8. The notion of irreversibility and the strength of an inscription is the subject matter of chapters 9 (a conceptual analysis) and 10 (an empirical case).

CHAPTER 7

Inscribing behaviour

Introduction

The two key notions of ANT from the previous chapter are inscription (which presupposes translation and program of action) and irreversibility (presupposing alignment). In this chapter we focus on the notion of inscription. Four aspects of inscriptions were identified in the previous chapter: the scenario, the material, who inscribes and the strength of an inscription. Inscriptions invite us to talk about how the various kinds of materials — artifacts, work routines, legal documents, prevailing norms and habits, written manuals, institutional and organizational arrangements and procedures — attempt to inscribe patterns of use (which may or may not succeed). Inscribing patterns of use is a way to confine the flexibility of use of an information infrastructure (see chapter 5).

So much for the underlying ideas of ANT and our minimal vocabulary. It remains to show how ANT may be employed to tease out important characteristics relevant to the design of information infrastructure. This chapter focuses on the notion of an inscription by illustrating what inscriptions in information infrastructures look like. They have many forms, quite a few of which are not easily spotted. We are accordingly particularly concerned with uncovering the different materials for inscriptions, that is, how and where patterns of use are inscribed. But first it is necessary to study how interests get translated, that is, how they are inscribed into one material before getting re-presented by inscribing it in a different material.

Focusing on inscriptions also means focusing on flexibility as the inscriptions of a technology are the links between the very technology and its use. There are two aspects of the flexibility of a technology: use and change flexibility respectively. Use flexibility means that the technology may be used in many different ways without changing the technology as such. This is its enabling character. Secondly, a technology may be more or less easy, i.e. flexible, to change in itself, when changes in requirements go beyond its use flexibility. Use flexibility means using the technology differently form what was intended, i.e. not following the inscribed program of action, while change flexibility means changing the technology according to a new programs of action.

All empirical material used in this chapter is from the establishment of an information infrastructure for EDI in health care in Norway.

Translations and design

Lab reports

The development of electronic information exchange between health care institutions in Norway started when a private lab, Dr. Fürst's Medisinske Laboratorium in Oslo, developed a system for lab report transmission to general practitioners in 1987. The interest of Dr. Fürst's laboratory was simply to make profit by attracting new customers. It was based on the assumption that the system would help general practitioners save much time otherwise spent on manual registering lab reports, and that the general practitioners would be find this attractive. Each general practitioner receives on average approximately 20 reports a day, which take quite some time to register manually in their medical record systems.

Fürst translated the interests of themselves as well as general practitioners (GPs) into a system. The system was very simple. It was implemented on top of a terminal emulator package, and the development time was only 3 weeks for one person.[1] Further, they enrolled the vendors of medical record systems for GPs into their network as well by paying them for adapting their systems to Fürst's module for receiving lab reports. The interests of GPs were further translated to enrol them into the network by giving them the modems needed to communicate electronically with Fürst for free.

1. Interview with Fiskerud (1996).

Lab orders

Labs also have an economical interest in receiving orders electronically as they could save a lot of labour intensive registration work. The ordering general practitioners, however, do not enjoy the same kind of immediate and tangible advantage. In order to enrol the general practitioners, electronic transmission of lab communication needs to translate the interests of the general practitioners' into the system. So far, this has not been resolved. Several attempts are being made, some of which will be presented here.

A crucial aspect of ordering tests is to ensure that an order and the specimen it belongs to are not mixed with others. A procedure followed by some general practitioners and labs today is the following: Each copy of the paper order is given a unique number. This number is printed on two different places on the form, including one adhesive label that is to be removed from the order and glued on the specimen container. In addition, the paper order is connected to the specimen container. Reproducing this level of security in the scenario when the order is transmitted electronically has turned out to be rather challenging, and will certainly include the design of specific technological as well as organisational arrangements. The design of a solution for lab orders invariably involves the alignment of the complete heterogeneous network of the collection of associated work routines as well as computer systems.

A possible solution that has been discussed is using a collection of label producing machines (bar code printers), label reading machines, manual routines and new computer applications. Each time an order is filled out, a bar code label will be printed by the general practitioner's system and subsequently glued to the specimen container. The unique number represented by the bar code is also a part of the specimen identifier in the order message. When the lab receives a specimen, a machine must read the bar code on the label and ensure that the specimen is attached to its proper order (already received electronically by the lab). The standardised message will inscribe the working routines for instance, the kind of information necessary for carrying out the control routines depends on how these routines are defined. However, as the general practitioners do not have any obvious advantages from electronic ordering, it is reasonably to expect that they are not interested in investing in bar code printers and other technological components these proposal demands.

During 1996, two different solutions have been tested out in Norway, each involving one lab and just a few general practitioners. One of them is based on what is called two dimensional bar codes. The complete order information is represented by a two dimensional bar code and printed on a label glued on the specimen container. The other solution is based on electronic transmission of the order using the standard European EDIFACT message, while a paper form is also printed and sent together with the specimen as in the current practice.

When orders are sent electronically, some new possibilities and advantages for the general practitioners as well are possible. One idea, which Dr. Fürst's lab wants to implement, is to take advantage of the possibility for ordering new tests of a specimen when the results of those ordered first are available. Usually a general practitioner orders several tests of the same specimen. Which combination of tests that are most interesting depends on the results of the analysis. Accordingly, it would be useful to order some tests, study the results and then decide on which additional tests that are relevant. When both orders and results are transmitted electronically, this possibility may become reality. It is, however, easier to implement this functionality using the on-line connection in Dr. Fürst's laboratory original, non-standardised solution. Such new services might be attractive to general practitioners and enable labs to enrol them into the networks necessary for making electronic ordering work. However, the programs of action inscribed into the standardised solutions based on EDIFACT and the ISO e-mail standards make it impossible to implement such services within that framework. Dr. Fürst's laboratory is interested in experimenting with communication technology to develop new or improved services for the general practitioners and their patients. They have so far judged EDIFACT technology too complex and inflexible and intends to wait until simpler and more flexible technology is accepted.[2] Web technology might fulfil the technical requirements for such technology, but this remains to be seen.

This example illustrates the wide range of different interests being involved, how some of them might be translated into aligned networks representing interesting and promising solutions while no actor have succeeded in building an aligned actor network strong enough to put the technology into real use.

Prescriptions

The idea of electronic transmission of prescriptions grew out of a feasibility study as part of Statskonsult's Infrastructure programme. This area was also identified as an interesting one in Telenor's Telemedicine project (Statskonsult 1992). Establishing an infrastructure for electronic exchange of prescriptions requires the alignment of a wide range of different interests, including general practitioners, pharmacies, patients, the government and social insurance offices.

2. Interview with IT director Sten Tore Fiskerud, Feb. 1996.

Unlike lab messages, there has up till now not been much done on an international level regarding electronic prescriptions. The effort in Norway we report on accordingly represents an early attempt to standardise prescription messages. As will become evident further below, the institutional arrangements of the standardisation process which link national and international efforts tightly, have resulted in a proposed, international standard for prescriptions heavily influenced by the Norwegian project.

The overall objectives of Statskonsult's Infrastructure programme was to improve productivity, service quality, and cost containment in the public sector. Spendings on pharmaceuticals are high, and accordingly an important area for cost containment. In addition, the health care authorities wanted enhanced control concerning the use of drugs by patients as well as prescription practices of physicians concerning habit-forming drugs.

The interests of the pharmacies were primarily improved logistics and eliminating unnecessary retyping of information (Statskonsult 1992). By integrating the system receiving prescriptions with the existing system for electronic ordering of drugs, the pharmacies would essentially have a just-in-time production scheme established. In addition, the pharmacies viewed it as an opportunity for improving the quality of service to their customers. A survey had documented that as much of 80% of their customers were favourable to reducing waiting time at the pharmacies as a result of electronic transmission of prescriptions (cited in Pedersen 1996).

As part of the Infrastructure programme KITH worked out a preliminary specification of an EDIFACT message for prescriptions (KITH 1992). The pharmacies also wanted to include information about so-called bonus arrangements (Norwegian: frikort) into this message. Certain categories of patients get (up till 100%) bonus on their drugs. This bonus is subsidised by the health insurance authorities on the basis of special reports from the pharmacies.

The interests of general practitioners in the project had different sources. Electronic prescriptions would eliminate retyping a lot of information which already was stored in the medical record system. It would also greatly support the reports the general practitioners send to the health insurance authorities, some of them being the basis for their payment. More importantly, however, electronic prescriptions were viewed as an element of the association of general practitioners' ongoing programme on quality assurance (cited in Pedersen 1996). Electronic prescriptions allow automatic cross-checking to be performed (for instance, that all fields are filled in properly). The general practitioners were also attracted by the prospects of getting access to the pharmacies' drug item list. This list is provided to the pharmacies by their provider of drugs through the pharmacies' application supplier (NAF-Data). The list contains information useful also for the general practitioners, for instance, about price and synonymous drugs. It is updated on a monthly basis. As we will spell out in more detail in the last section of this chapter, this list turned out to become the source of much controversy.

A key challenge in the prescription project was to find a way to align the interests of the involved actors, most importantly the pharmacies and the general practitioners. According to actor network theory, this takes place by translating these interests and inscribing them into a material. This drug item list play the role of such a material. Today, the list of drugs accessible to the general practitioners medical record system is either manually entered and updated or is provided through the vendors of medical records systems at a substantial cost.

The failure of building standardized infrastructure

The development and adoption of the network designed by Fürst was a smooth and effective process and the solution has been very useful. The same was more or less the case with the copied solutions installed by other labs. However, later efforts aiming at developing similar solutions for other areas have failed. Important explanatory factors, in terms of actor network theory, are partly the fact that the prime actors have not managed to translate the interests of those they have delegated roles in their design into the very same design in a way making the whole actor network aligned. In particular they have failed in translating the rather general design of their EDI systems into the working situations of its anticipated users.

Another important factor is the fact that the general, universal, solutions arrived at generates a very large and complex actor network, so complex that it cannot be aligned in any proper way in a changing world.

Inscriptions and materials

It is close to a cliche to maintain that technology, including information infrastructure, is never neutral, that certain patterns of use are encouraged while others are discouraged. By studying the material of inscriptions, this may be pushed further by specifying more specifically how and where behaviour is (attempted) inscribed. The variety of the material of inscriptions is indicated in a top-down fashion by illustrating inscription on a high, organisational level, on an architectural level, in the messages as well as in tiny, grey details contained within a specific field in the message.

Inscribed in the bureaucratic organisation

The Norwegian efforts at developing communication between labs and ordering hospitals and general practitioners were quickly and tightly aligned with international efforts, especially those by CEN, the European branch of ISO. Translating what started as a relatively down-to-earth, practical endeavour in Norway into a European arena inscribed unintended behaviour and consequences. From the very beginning, there was a fairly close dialogue between the Norwegian designers of lab communication and their users. When the design was translated into a European effort to promote the European Union's visions for an open market, the users were unable to follow it. The problem had been moved to an arena with many and unknown bureaucratic rules of the game, circulation of technical descriptions rather than practical problems and heaps of paper rather than operative solutions. It is time to take a closer look at the inscriptions of the EDIFACT based bureaucracy. To be involved in the CEN standardization, the ongoing Norwegian effort had to be translated and aligned with this bureacracy.

EDIFACT is not a self-contained piece of technology. It is a heterogeneous actor-network which includes: syntax for defining data structures; tools like converters and data bases for definitions of messages and message elements; a hierarchy of standardisation bodies on global, regional (i.e. European, American, etc.) and national levels; prevailing conceptions and established practices for how to define and implement messages; an EDIFACT industry of vendors and consultants; artifacts like manuals and other forms of documentation, and educational material about EDIFACT.

The size and complexity of this network make its inscriptions strong and difficult to work against when one is enrolled into it. We will first look at programs of action related to the standardisation process of EDIFACT, then we turn to patterns of use inscribed in the EDIFACT technology itself.

EDIFACT technology and the organisation of EDIFACT standardisation processes make it virtually impossible for users to be involved in, not to say influence, the standards setting processes. The standards are controlled by a group of more or less professional standardisation people who work for large companies or bureaucracies. Inspired by MacKenzie's (1990) notion of the "gyro mafia", this group may be dubbed the "EDIFACT mafia". This mafia's control is neither a feature of the EDIFACT format itself nor the organisation of the standardisation process, but it is a result of the interplay between the elements of the EDIFACT actor-network outlined above.

An unintended consequence of the complexity and non-transparency of the EDIFACT actor-network is that it inscribes barriers on end-user involvement through its requirements on the level of competence. To be involved in the standardisation work, one needs to know all the rules of the game - the technological details of EDIFACT, the formal rules of the standardisation bodies as well as all the informal practices. There are formal and informal rules for how a message should be composed as well as how the processes should run. An essential EDIFACT rule is that existing standardised messages and message elements should be used as far as possible when defining new ones. This implies that in order to make lab standards, one also has to be familiar with standards within virtually all other sectors as well. The effect, unanticipated we assume, is that it preserves and professionalises the mafia's control over the process.

In addition, the tendency within EDIFACT to emphasise the technical aspects delegates an even less central role to end-users. The specification of the data format used in the first proprietary systems literally fits on one page of paper and is easily understood by those who need it. The specification of the European standardised EDIFACT message, however, is a voluminous document of 500 (!) pages (ref CEN 1994a, 1994b). Where this message is used, the information exchanged is almost exactly the same as when using the old systems (CEN 1992b, 1993a, 1993b; KITH 1994)! The bias in lab communication standardisation towards the technical and general issues at the expense of the practical is shared by other EDIFACT efforts as documented by the evaluation of the European Union's programme on diffusion of EDI in the trade sector, the TEDIS programme (Graham et al. 1996). In this sense, the bureaucratic and procedural arrangements of EDIFACT inscribe few and only indirect opportunities for user influence.

Inscriptions in the systems architecture

A systems architecture is its overall, organising principle. This is a somewhat vague notion. In our case it can be used to distinguish among architectural categories like transaction oriented solutions, event driven ones, message oriented systems, client-server oriented ones, etc.

The choice of such a systems architecture is not neutral. It is the material for inscriptions. We will illustrate this by the looking more closely at the inscriptions of the EDIFACT bias towards a message oriented systems architectures.

EDIFACT inscribes certain patterns of use. This is partly inscribed in the broadly established view that EDIFACT is mimicking today's physical exchange of paper forms, orders and invoices being paradigm examples. This view is also translated into the choice of communication carrier for exchanging EDIFACT messages, i.e. using e-mail as standardised by the ISO. Using e-mail implies that the receivers get information when the senders want to provide them and not when receivers themselves want it. For clinical-chemical laboratories, for instance, the results will be sent to the ordering general practitioner when the ordered tests are completely analysed, or at predefined intermediate points in the analysis process. This inscribes a behaviour which blocks what is possible with some existing, non-standardised systems. The Fürst laboratory in Norway and its customers use a system where the general practitioners at any time may access the results produced in the analysis processes up to that very moment in time. This function will not be provided by the standardised, e-mail based solution. Other EDIFACT inscriptions will be touched upon in later sections.

Inscriptions in the message syntax

Compared to modern programming language constructs, the EDIFACT syntax - or data definition mechanisms - is quite primitive. These shortcomings inscribe centralised control and barriers to flexible appropriation to local contexts of use (Hanseth, Thoresen and Winner 1993). The non-transparency of the overall EDIFACT actor-network tends to make these inscriptions invisible and hence unanticipated.

Technically speaking, the EDIFACT syntax lacks constructs for subtyping (or inheritance), pointers and recursive data structures. The ability to subtype would come in very handy when defining standards covering different geographical areas and different disciplines. Subtyping provides a mechanism for defining a standard as a collection of modular building blocks. The lab messages have been defined in order to serve the purpose of a large number of labs (for instance, clinical-chemical labs, micro-biological labs and X-ray labs). In addition, there are geographical differences. Using EDIFACT, a number of different subsets or specialisations of the message have to be worked out. As support for subtyping is lacking, the only way of enabling this is to define a European message covering *all* local variations as optional elements. Local specialisations are then defined by specifying which of the optional elements are mandatory and which ones should not be used. With subtyping, local modifications would be contained within one module, leaving all other unchanged.

In this way the EDIFACT syntax inscribe centralized control of the standards and standardization process, inhibiting user participation, etc. The syntax is well aligned with the bureaucratic organization, embodying the same inscriptions. Together they make these inscriptions stronger.

Inscriptions in the message itself

An important part of the definition of standardised messages is deciding which data elements should be included in the messages and which should not. These elements are also material for inscriptions.

In the system Dr. Fürst laboratory developed, only basic result data were included. The Health Level 7 message used later on as a prototype, included more information. Reflecting the organisation of the health sector in the United States with private financing, economic information was included. Some economic information may be relevant in Norway as well, especially if the message is seen in the context of the overall economic organisation of the sector, that is, who is paying for what, who is responsible for quality control and cost containment, which institutions are involved in the payment and what kind of information they need.

Based on use scenarios worked out in the Norwegian lab messages working group during 1991-92, it was concluded that the data set in the Health Level 7 message did not satisfy the needs (KITH 1991). The message proposal was distributed together with a request for comments. It was, however, decided that economic information should not be included in the first official message standard for reasons of simplicity. This was controversial. The association of pharmacies, for instance, expressed in their comments that the areas of use should be expanded to include information exchange between labs, general practitioners and institutions outside health care such as social insurance and banks.

In some European countries, the patients (through the general practitioners) pay part of the costs of the tests, but not in Norway. For this reason, the price the general practitioners pay for each test is included in the European report message. The general practitioners are invoiced periodically. The price information is important in order to control that the amount they have invoiced is correct. Accordingly, the European standard message include this economic information, and so does the Norwegian subset.

Another open issue was whether the information in a lab order should be included in the result message as well. Usually the result is returned to the ordering physician knowing the order specification already. Accordingly, in most cases the order information would be unnecessary. In some situations, however, the result is returned to another general practitioner than the ordering one. This is the case in ambulatory care, where the general practitioner visiting the patient orders a test while the result should be returned to the patient's ordinary general practitioner. In hospitals the ordering general practitioner may have left work and a new one has taken over the responsibility for the patient when the result arrives. In these cases, the result should include the order information as well. If this information is not available, the general practitioner may try to guess (which in many cases would work pretty well), or call the lab and ask them.

The arguments against including the order information are the increasing complexity and size of the messages and message specifications it leads to. One proposal put forth was to send the order as a separate message when needed. This solution needed a reference in the result message to its corresponding order message to avoid confusion. Such references, however, are not a part of EDIFACT as it is used. Technically, it would be very simple to find a working solution. The problem was that it would not follow the “rules of the game” of defining EDIFACT messages. It worked against deeply inscribed practises of specific ways to use EDIFACT. Accordingly it was ruled out. It was instead decided that the order information could be included in the result message.

These examples illustrate that the inclusion or not of a data element in a message is an negotiation over which programs of action should or should not be inscribed into the standard. In these negotiations, EDIFACT acts as a powerful actor in the sense that most alternatives are close to the intended and customary way of using EDIFACT.

Inscribed into a field in the message

EDI based information infrastructures usually need various identification services. Lab report messages must identify its order, the ordering unit (GP, hospital department, another lab, etc.), the patient, etc. Prescription messages must identify the prescribed drug. Registers of such identifiers must be available to the II users. This requires an II for maintaining and distributing updated versions of the registers. Such IIs may be rather complex technologically as well as organizationally. At the European level, there has been much discussion about a European standardized system of codes for identifying the object analysed (blood, skin, urine, etc.), where on the body it is taken from, the test performed and its result. It has so far turned out to be too difficult to reach an agreement. The problem is partly to find a system serving all needs. But it is maybe more difficult to find a new system which may replace the installed base of existing ones at reasonable costs. For more information on these issues see (Hanseth and Monteiro 1996), and (Bowker and Star, 1994).

Another example of inscriptions into single elements is given in the following section.

Accumulating the strength of an inscription

We have described how actors seek to inscribe their interest into heterogeneous materials. Through processes of translation, these interests may be inscribed into a variety of materials. There are more than one way to get the job done.

But inscriptions are often unsuccessful in the sense that their scenarios are not followed, their intentions not fulfilled. To increase the likelihood that an inscription may succeed, it is necessary to increase its strength. A key insight is that, in priciple, it is impossible to know beforehand whether an inscription is strong enough to actually work — it remains an open, emperical question that needs to be addresses by a strategy of trials and errors.

The question, then, is how do you increase the strength of an inscription? We illustrate two strategies. The first is the superimposing of inscriptions. Rather than merely observing that one and the same scenario may be inscribed into this material or translated into that one, you “add” the inscriptions. Instead of an either-or kind of logic you adopt the both this-and-that kind of Winnie the Poe logic. The second strategy is to expand the network (in ANT terms, enroll some actors&technologies) and look for new, as yet unused, material to faithfully inscribe your scenario into.

According to actor network theory, inscribing behaviour into actor networks is how an actor might reach an objective. Two forms of objectives are:

1. To make necessary support to make a favorable decision and implement it.
2. To succeed in the design and implementation of a system.

In general, building alliances is a useful strategy for realizing one’s will. This is one of the main ideas behind actor network theory. An alliance is built by enrolling allies into an aligned network. However, to obtain both objectives mentioned above, the network of allies will include humans as well as technologies (REFXX-Latour, “The Prince”). To get support for a decision, you have to trestle technologies to fit the whole alliance you are building. When designing technology, you are also designing roles for humans, for instance users, support people, etc. To make the technology work, you have to make them play the role you have designed.

The two examples in this section illustrate how inscriptions are made stronger through both these two ways of alliance building.

Technology as ally

As there was a general consensus — the interests were aligned — about the need for standards, the fight about what these standards should look like and how they should be developed started. This race was a seemingly neutral and technical discussion about which technology fitted the needs best. In reality, however, it was a race between different actors trying to manoeuvre themselves into key positions as “gatekeepers” or “obligatory passage points” (Latour 1987). In this race, most of them chose the same generic strategy, namely to first look for the technology which seemed most beneficial for them and subsequently enrolling this technology into their own actor-network as an ally. Appealing to the symbolic character of, technology makes it possible to make non-technical interests appear as technical arguments. We will here present some actors and how they were selecting technologies as allies or strategic partners. The general strategy was first to find an appropriate technology which each actor was well “equipped” to represent, and second, making the allied technology a powerful actor by socially constructing it as the best choise as standard.

Based on their interests in general solutions and rooted in the telecommunication tradition of international standardisation, Telenor searched for international activities aiming at developing “open” standards. The IEEE (Institute of Electrical and Electronics Engineers) P1157 committee, usually called Medix, did exactly this. This work was the result of an initiative to develop open, international standards taken at the MEDINFO conference in 1986. Medix, which was dominated by IT professionals working in large companies like Hewlett Packard and Telenor and some standardisation specialists working for health care authorities, adopted the dominating approach at that time, namely that standards should be as open, general and universal as possible.

The appeal for open, universal standards inscribed in the Medix effort implied using existing OSI (Open Systems Interconnection) protocols defined by the ISO (International Standardisation Organisation) as underlying basis. The Medix effort adopted a standardisation approach — perfectly in line with texts books in information systems development — that the development should be based on an information model being a “true” description of the relevant part of reality, that is, the health care sector, independent of existing as well as future technology. This case will be presented in more detail in the next chapter Individual messages would be derived from the model more or less automatically.

While the focus was directed towards a comprehensive information model, lab reports were still the single most important area. However, for those involved in Medix the task of developing a Norwegian standardised lab report message had around 1990 been translated into the development of a proper object-oriented data model of the world-wide health care sector.

In addition to the information model, protocols and formats to be used had to be specified. In line with the general strategy, as few and general as possible protocols and formats should be included. Medix first focused on Open Document Architecture believing it covered all needs for document like information. However, around 1990 most agreed that EDIFACT should be included as well. The Europeans who strongest advocated EDIFACT had already established a new body, EMEDI (European Medical EDI), to promote EDIFACT in the health sector. In Norway, a driving force behind the EDIFACT movement was the "Infrastructure programme" run by a governmental agency (Statskonsult) during 1989 - 92. Promoting Open Systems Interconnection standards and EDIFACT systems based on Open Systems Interconnection were key goals for the whole public sector (Statskonsult 1992).

The Norwegian branch of Andersen Consulting, pursuing clear economical interests, was marketing a product manufactured in the United States based on the so-called Health Level 7 standard. To promote their product, they pushed Health Level 7 as a standard in Norway even though it was evident that a substantial modification to make it fit a Norwegian context was required.

A second vendor, Fearnley Data, decided during 1989 to develop products supporting information exchange within the health care sector. They followed the Medix as well as the Health Level 7 activities. In early 1990, they initiated activities aiming at developing Health Level 7 based Norwegian standards. They organised a series of meetings and tried to enrol the necessary actors into a network aligned around Health Level 7 with themselves as the main gatekeeper while at the same time keeping Andersen Consulting outside the network by focusing on the amount of work required to modify their product in Norway.

In 1990 the Ministry of Health decided that standards should be developed. Responsible for this work was Gudleik Alvik. He hired a consultant, Bjørn Brevik, for doing the technical work. He specified a coherent set of data structures and exchange formats along the same line as that of Dr. Fürst's and similar systems (ref.). The proposal was distributed for comments. The procedure followed by the ministry was the general one used for all kind of decision making concerning new rules to be followed by health care institutions. This procedure was — of course — very different from those of international telecommunication standardisation. It delegated power and competencies to actors within the health care sector and not the telecommunication world. Actors from the telecommunication world mobilised and easily killed this proposal (ref).[3]

KITH was established in 1991 and was delegated the responsibility for standardisation by the Ministry of Health. KITH's director Bjørn Engum was the former head of Telenor's telemedicine project, and accordingly enrolled into their standardisation activities. He aligned closely with Statskonsult and likewisely argued in favour of EDIFACT and OSI. As was the case with Telenor's role, key public institutions made heavy use of their perceived neutrality to advocate their interests.

Late in 1990, Fearnley Data started the development of a communication system for health care. At this time they had given up the Health Level 7 based standardisation approach because EDIFACT was gaining momentum. They decided to ally with EDIFACT rather than Health Level 7. They furthermore aligned with other standardisation bodies and activities, including European Medical EDI, KITH and Statskonsult. At the same time, another company (Profdoc) started the development of a product paying less attention to standardisation and rather more to the experiences with existing systems.

Fearnley Data decided that their product should follow standards as far as possible. When they started, no formal decision about Norwegian or international standards had been made. However, a "rough consensus"[4] had been reached that EDIFACT should be the basis for the exchange of structured form-like information. Accordingly, Fearnley Data considered it safe to start the implementation of an EDIFACT based solution. One of their employees, Edgar Glück (educated as a doctor, practising as a systems designer) designed a first version of a lab report message in EDIFACT based on the Health Level 7 message. Fearnley Data's strategy was to install their solutions for communication between hospital labs and general practitioners' offices in parallel with promoting the message as a proposed standard within national and international bodies. This strategy turned out to be very successful. The existence of a specified message and "running code" had similar effects as Dr. Fürst's system. As Fearnley Data had one of the very rare existing EDIFACT implementations, they were highly successful in mobilising support for their solution. Having a concrete solution, as opposed to merely playing with paper sketches, proved to be an effective way of enrolling others. With minor changes the message was accepted by both KITH and EMEDI. EMEDI sent the message specification to the Western European EDIFACT Board as a proposal for a message being formally

3. One of the authors was involved in this killing.

4. The Internet slogan "We believe in rough consensus and running code" is indeed a precise description of their successful strategy (Hanseth, Monteiro, Hatling 1996). This slogan nicely captures the successful strategy behind the development of the Internet.

approved by the international EDIFACT standardisation authorities. The message was quickly approved.

The alignment of interests and technologies around EDIFACT established a very powerful actor-network. EDIFACT technology in general and running EDIFACT solutions in particular were powerful actors in this alliance. Profdoc reported that it was “impossible” from 1992 to market their product as it was not based on EDIFACT and standards from the ISO. The rhetoric of “open” standards was quite effective.

In 1990 the Commission of the European Community delegated to CEN (Comite Europeen de Normalisation, the European branch of ISO) to take responsibility for working out European standards within the health care domain in order to facilitate the economical benefits of an European inner market. CEN established a so-called technical committee (TC 251) on the 23. of March 1990 dedicated to the development of standards within health care informatics. From this time Medix disappeared from the European scene. However, the people involved moved to CEN and CEN’s work to a large extent continued along the lines of Medix.

When CEN started their work on lab reports, some proposals existed already. For this reason, they wanted to build upon one of these (CEN 1991). They hoped the message specification worked out by Edgar Glück and approved by European Medical EDI could be proposed as a pre-standard. If so, a pre-standard for lab information could be ready already in April 1992. There was a great pressure for producing results rapidly. However, groups allied with other technologies than EDIFACT opposed this. Among these was a group consisting of just a few persons being involved in the Euclides project under the first, preparatory phase of the European Union’s health care telematics programme.

The Euclides project developed a prototype of a system for lab report exchange based on their own non-standard format. After the project was completed, a company was set up in Belgium to continue the work. Being a European Union project, Euclides was well known in the European networks which the CEN work was a part of. As the CEN work was financed by the European Union, the Euclides project was perceived as more important and relevant than its size and achievements would imply. An additional, important factor was the fact that the head of the health care committee of CEN (TC 251), George De Moor, was also the manager of the Euclides project.

The Euclides group realised that they would not succeed in trying to make their format and message the only European standard. Accordingly, they made an alliance with the information modelling approach, proposing to develop an information model for lab first, and that this model would be the primary standard. Based on this model the information could be exchanged using EDIFACT as well as other formats. This proposal was inherited from earlier Medix work, channelled to CEN by former Medix people. As more countries participated in the health care committee of CEN (TC 251) than the EMEDI group, it was decided to adopt the information modelling approach instead of modifying the EDIFACT message approved by EMEDI. This work was extended by a specification for how information should be exchanged using EDIFACT. To our knowledge, how to exchange the information using Euclides or other messages or formats have not been specified

In this process of searching for technologies as powerful allies, these were found among the general and well established ones. As EDIFACT was gaining momentum in Norway as well as Europe at this time (early 90s), EDIFACT — together with the most closely aligned standardisation bodies — did occupy centre stage. The strategy first adopted by Dr. Fürst's laboratory (accumulating practical experience from various contexts of use within the health care sector) was abandoned in favour of a strategy focusing on modelling techniques. This did not inscribe definite programs of action, but it did inscribe a shift in the delegation of competence about health care to competence in software engineering. This delay of gaining practical experience by aligning with international standardisation bodies inscribed fewer and less direct channels for end-user input from the health care sector.

Expand the network to accumulating strength

In chapter 6 we explained how, according to actor network theory, inscriptions have to be linked to larger actor-networks in order to give them sufficient strength. Exactly what it takes to make an inscription strong enough is not possible to know beforehand, it is a question of practical trial and error. A program of action is inscribed into an increasingly larger actor-network until the necessary strength is reached. This aspect of actor network theory is nicely illustrated, we believe, by the attempts presented below to inscribe a desired behaviour of general practitioners into the definition of the semantics of one single data element in the prescription message. The question, then, is how to accumulate enough strength for this inscription to actually enforce the desired behaviour of general practitioners. Most examples presented above in a similar way illustrate how EDIFACT inscriptions have accumulated its considerable strength.

A principal reason for the interest in prescriptions from the point of view of the pharmacies was the prospect of improved logistics by integrating the existing electronic ordering of drugs from the drug depot (Norwegian: Norsk Medisinal Depot) (Statskonsult 1992; KITH 1993a). To exploit the economically interesting possibilities of establishing this kind of just-in-time distribution scheme, there had to be a way for the pharmacies to uniquely identify a prescribed drug with the drug which subsequently was ordered from the drug depot. In the electronic ordering of drugs from the drug depot, the pharmacies made use of an existing drug list with a coding scheme for drug identifiers as a six digit article number.[5] This drug list was updated and maintained by the drug depot.

The pharmacies' interests for improved logistics was accordingly translated into a proposal to include this six digit drug identifier into the electronic prescription message. This way of inscribing their interests into the semantics of one data element in the message was proposed by the representative of the pharmacies early in the pre-project (KITH 1992)[6].

No one seems to have objected to this proposal from the pharmacies despite (or may be because of) the fact that the scenario of use which was inscribed was not spelled out in any detail. In particular, the pre-project did not spell out exactly *how* the general practitioners should provide this drug identification number when making a prescription. The general practitioners do not make any use of this number. They identify drugs by their type or brand names, not their identification number. It is not feasible to increase the workload of general practitioners by demanding that they provide the identifier manually. In that case, electronic transmission would require more work than the paper prescriptions and the general practitioners would have no incentives to change.

Rather than working out a detailed program of action, the representative from the general practitioners' associations suggested that this somehow could be solved if the general practitioners were granted access to the list of drug identifiers the pharmacies had, the list maintained by the drug depot. Gaining access to this list was appealing to the general practitioners for two reasons. Besides the drug identifiers, the list contains other information useful for the general practitioners such as prices and synonymous drugs. The fact that the list is continuously updated was also considered favourable. When the pre-project ended in 1992, what remained was to translate the general practitioners' interests in accessing the drug list into a suitable (but unspecified) inscription and align this with the already agreed upon inscriptions in the prescription message. In actor network theory terms, the inscription

5. Reaching agreement on article numbers has been an important and challenging part of the establishment of EDIFACT networks in several business sectors.

6. This should not be taken to imply that the pharmacies had their ways in every respect. At the same meeting the pharmacies also suggested including a data segment for bonus arrangements which would have substantially improved their reporting routines the health insurance authorities. This suggestion was declined, mainly for reasons of simplicity (KITH 1992).

which demanded that general practitioners provide the drug identifier was to be strengthened by aligning it with a larger (but unknown) actor-network inscribing access to the drug list for general practitioners.

The proposals from the pre-project (KITH 1992) were circulated for comments. Profdoc, the sceptical vendor of electronic medical record systems (see above), was also critical to how the issue of drug identification numbers should be solved (Profdoc 1993). The solution Profdoc suggested was to extract the identification number from another source, the so-called “Common Catalogue” (Norwegian: Felleskatalogen) instead of the pharmacies’ drug list. The “Common Catalogue” is a paper based catalogue which all general practitioners have. It contains information about all registered drugs in Norway including their identification number. In addition, it contains information about treatment of acute poisoning, drugs that interfere each other, and a register over drug producers and pharmacies in Norway. The catalogue is printed once a year, while additions regarding new or obsolete drugs are printed and distributed continuously. The “Common Catalogue” is produces by a publisher (Fabritius) and was recently available also electronically in the form of a CD-ROM. This solution based on the “Common Catalogue” delegates a very different set of roles to the involved actors. The required integration work between the electronic medical record system and the prescription module would now involve the publisher but neither the pharmacies nor the drug depot. Besides simply pointing out a, technically speaking, perfectly feasible alternative to a solution based on the drug list from the pharmacies, Profdoc also had a more self-centred interest in promoting it. During the period after the pre-project was completed, Profdoc had a series of meetings with the publisher of the “Common Catalogue.” Profdoc explored the possibility, independently of the prescription project, to integrate their medical record system with the “Common Catalogue.” They had never taken pro-active part in the prescription project. When the issue of drug identification number surfaced, they apparently seized the opportunity of trying to design a solution delegating a role for themselves and their allies in the prescription project.

The alternative suggested by Profdoc was not pursued in the main project. Instead, the project continued to work on how to make the drug list available. This soon turned out to be a lot more complicated than they imagined. The heart of the matter was that the list belonged to an actor outside the project, namely the drug depot. As the list contained information which was confidential, for instance about profit margins on pharmaceutical products, the drug depot had commercial interests in it and refused to hand it over free of charge. Hence, the attempts to accumulate strength for the inscriptions demanding that general practitioners provide the drug identifier were faced with serious, unforeseen problems. It was necessary to translate the commercial interests of the drug depot, a non-project actor, into an inscription. This would involve inscribing roles and obligations for (at least) the following issues: how to obtain the list from the drug depot, how to "wash" the list to make it appropriate for use for general practitioners, who should do — and pay for — the work. The fact that the participants in the project had to finance their activities themselves, made negotiations difficult. The problems with working out an agreement with the drug depot dragged on. In a coordination meeting in January 1994 it was stated that an agreement was to be reached.

Late in 1995, the testing of the system for electronic transmission of prescriptions started at a pilot site (one general practitioner and one pharmacy). In this first version of the system, drugs are identified by their ordinary brand names. Employees at the pharmacy will map this name to its identification number manually. When the name is incomplete or misspelled, as it is assumed quite often will be the case, they will call the general practitioner by telephone. This version will not be used for reiterated prescriptions either.

Due to the European Economical Area treaty, the earlier monopoly status of the drug depot has been dismantled as of 1. of January 1995. This paved the road for several distributors of drugs to pharmacies beside the drug depot. Each would have their own drug identification number scheme as no "global" identification coding scheme exists. This makes the drug depot's earlier situation a lot more vulnerable. To the project leader, the drug depot has stated that they now are willing to let give general practitioners free access to their drug list (Yang 1995). During 1996, the provider of applications for the pharmacies, NAF-Data, has been setting up a data base for all distributors of drugs in Norway including the drug depot. This data base is intended to be made accessible to the general practitioners. It has been decided that a new institution will be established and delegated the responsibility for giving each drug its unique identification number.

Moving on

The notion of an inscription is, as we have argued both analytically as well by drawing upon empirical material, a fruitful and instructive notion when analysing information infrastructures. It may be employed in order to underscore the (potential lack of) flexibility in an information infrastructure. In the chapter that follows we analyse the contents of and the background for the lack of flexibility in many of the information infrastructure initiatives. They are based on an (ill-conceived) assumption that the existing, heterogenous bits and pieces of an infrastructure do not inscribe any behaviour, that it is more or less straightforward to sweep alternatives aside and establish a new, universal solution for covering all patterns of use.

In the following chapter 9, we explore further the way inscriptions accumulate strength by “adding” or superimposing one on top of another in the way we have explained in this and the previous chapters. The total effect of such an “adding” and aligning of inscriptions is irreveribility. This process is particular relevant to come to grips with when analysing information infrastructures due to the long time it takes to establish an infrastructure. It is never established from scratch, only by gradually expanding, substituting and superimposing new elements. This entails that the actor-network in total — the installed base of existing components and patterns of use — inscribes and hence influences the further development and use of the information infrastructure.

CHAPTER 8

Dreaming about the universal

Introduction

Most information infrastructure standardization work is based on a set of beliefs and assumptions about what a good standard is. These beliefs are strong — but not based on any empirical evidence concerning their soundness. They have strong implications for what kinds of standards that are defined, their characteristics as well as choice of strategies for developing them. Beliefs of this kind are often called ideologies. Hence, the focus of this chapter is on the dominant standardization ideology which we dub “universalism”: its content, history, how it is tried applied, what really happens and its shortcomings. We will argue that it has serious shortcomings. In fact, dominant standardization approaches do not work for the development of future information infrastructures. New approaches based on different ideologies must be followed to succeed in the implementation of the envisioned networks. Basically, the heterogeneous nature of information infrastructures pointed out in chapter 3 needs to be acknowledged. Furthermore, the way the different components enable and hamper flexibility through their inscriptions needs to be emphasised a lot stronger. To this end, we discuss how visions and patterns of use and strategies for change are inscribed into standardisation efforts dominated by the ideology of universalism.

What is universalism?

Universalism is the dream about the perfect, “universal solution,” the “seamless web” where everything fits together perfectly and where any information may be exchanged between anybody connected without any loss of meaning or other obstacles. This seamless web is believed to be realized through the implementation of consistent and non-redundant standards. Dreaming about such a technology seems perfectly sound. This list seems like innocent and desirable characteristics of any well designed information system, including information infrastructures. This is, of course, true in a sense. This is a goal to strive for in most design efforts, including design related to IIs. But striving for this goal will often also cause serious trouble. Exactly because universalism is immediately appealing, it is an extremely strong rhetorical device, and correspondingly difficult to see what is wrong with it. After all, who would rather have a messy, complex solution than a elegant and coherent one? It does not, after all, make much sense to invent a lot of wheels in parallel. A much better idea is to just make one and let everybody use it.

There is only one major drawback with this scheme: it does not work just this way. Universal solutions of this kind presuppose an overarching design comprising the information infrastructure as a whole which is not attainable. It is based on a “closed world” set of assumption. As the world of infrastructures is an open one without limits, the universalism implies trying to make a complete, consistent formal specification of the whole unlimited world - which is simply impossible. Universalism also implies homogeneity as opposed to the heterogeneous character of information infrastructures emphasized in this book (see chapter 3).

In practice universalism leads to big, complex, incomprehensible and unmanageable standards and infrastructures. There cannot be any tidy, overarching design. It is a bricolage of components sometimes developed for different purposes, at different times. Hence, the inclination towards universal solutions, although understandable, needs closer scrutiny. Its benefits are greatly exaggerated and its problems vastly down-played (Graham et al. 1996; UN 1996; Williams 1997).

Universalism is not unique to standardization work. In fact, it is a strong ideal for virtually all technical and scientific work. In this chapter we will look at how universalism is imprinted on health care standardization work in particular and other information infrastructures and technical and scientific work in general. We will further look at the arguments given for the ideals of universalism, how they are tried implemented, what really happens when following the ideology. Finally we will analyse the experiences and identify its shortcomings.

Universalism at large: a few illustrations

Universalism, is expressed in a variety of ways and in numerous situations. Our ambition is neither to be systematic nor comprehensive, but merely to provide enough illustrations to make our point, namely that the bulk of standardisation has been dominated by this ideology.

Health care

The CEN TC/251 clearly expresses universalism as its dominant ideology. Hence, despite the fact that an information infrastructure will evolve, we have to pay the price of "freezing" it into one, given, universal solution:

> "in case of moving technologies ... standards could possibly impede the development. On the other hand, it may be desirable to make sure that unsuitable circumstances (e.g. proliferation of incompatible solutions) are not allowed to take root and in that case, standardization must be started as soon as possible in order to set the development on the right track" (De Moor 1993, p. 4).

The assumed need for a coherent, consistent and non-redundant set of global standards are even more clearly expressed by HISPP,[1] the US coordination committee collaborating closely with CEN. The great fear of universalism, fragmentation and hence mess, is warned against:

> "the efforts of these groups have been somewhat fragmented and redundant. Parallel efforts in Europe and the Pacific Rim threaten to further fragment standardization efforts" (McDonald 1993).

> "If we could eliminate the overlap and differences, we could greatly magnify the utility and accelerate the usage of these standards." (McDonald 1993).

The danger, it is maintained, is that of barriers to the free flow of commodities and services:

1. HISPP (...) is coordinating the ongoing standardization activities in different standardization bodies in US.

> "The establishment of a 'Fortress Europe' and the creations of barriers of trade have to be avoided" (De Moor 1993, p. 6).

Universalism goes deep. In the very statement of principles for the work of the health care standardization, CEN TC 251, it is explicitly pointed out that redundancy is not acceptable because "duplication of work must be avoided" (De Moor 1993, p. 6).

Similar concerns are voiced by the work coordinated in the US by HISSP which hold that there is:

> "too much emphasis on short term benefits while ignoring the long term benefits that would come through the analysis of large data bases collected in uniform fashion over large patient populations" (McDonald 1993).

The belief in one universal standard is explicitly expressed:

> "[The goal is] achieving ... a *unified set of non-redundant, non-conflicting* standard" (McDonald 1993, p. 16, emphasis added).

This is a quite strong expression as it refers to the work of all groups developing health care information infrastructure standards in US as well as in the rest of the world and standards outside the health care sector like EDI standards in general.

Striving for maximally applicable solutions in a world without clear borders has its implications. As different use areas are linked together and are overlapping, developing coherent solutions for subfields, implies making a coherent solution for everything. The same inscriptions and patterns of use are accordingly assumed to equally reasonable everywhere.

The difficulties in defining strict borders and hence the need for all-encompassing solutions were acknowledged in the Medix effort. The objective of this effort was defined as the development of one single global standard, or one coherent set of standards, covering any need for information exchange within health care:

> "The eventual scope of P1157 is all of healthcare communications, both in the medical centre, between medical centres, and between individual providers and medical centres" (ibid.).

This was carried over to CEN. CEN consider standards the key to successful health care II development, and that standards should be developed at an international, hopefully global, level:

> "More alarming is the ascertainment that many of these telematics-experiments [in US, Europe (typically projects sponsored by EU), ..] in Health Care have been conducted on regional scales [i.e. Europe, US, ..] and inevitably have resulted in the proliferation of incompatible solutions, hence the importance of standardization now" (De Moor 1993, p. 1).

which implies that

> "Consensus standards are needed urgently" (De Moor 1993, p. 2).

Computer communications

Universalism has an equally strong position within telecommunication and computer communication in general. The OSI model and the OSI protocols are clear examples of this. The protocols are defined under the assumption that they will be accepted by everybody, and accordingly the only ones in use. Not addressing how they should interoperate with already existing network protocols, i.e. their installed base hostility," has been put forth as the major explanatory factor behind their lack of adoption.

The Internet has often been presented as the opposite alternative of OSI. That might be done here as well (Abbate 1995).

Internet was based on the idea that there would be multiple independent networks of rather arbitrary design, beginning with the ARPANET as the pioneering packet switching network, but soon to include packet satellite networks, ground-based packet radio networks and other networks. The Internet as we now know it embodies a key underlying technical idea, namely that of open architecture networking. In this approach, the choice of any individual network technology was not dictated by a particular network architecture but rather could be selected freely by a provider and made to interwork with the other networks through a meta-level "Internetworking Architecture".

In an open-architecture network, the individual networks may be separately designed and developed and each may have its own unique interface which it may offer to users and/or other providers, including other Internet providers. Each network can be designed in accordance with the specific environment and user requirements of that network. There are generally no constraints on the types of network that can be included or on their geographic scope, although certain pragmatic considerations will dictate what makes sense to offer.

In this way the Internet was developed under the assumption that there would be no one single, universal network, but rather a heterogeneous infrastructure composed of any number of networks and network technologies. These assumptions are even more explicitly expressed in the four ground rules being critical to Robert Kahn's early thinking, which later on led to the design of TCP/IP:

- Each distinct network would have to stand on its own and no internal changes could be required to any such network to connect it to the Internet.
- Communications would be on a best effort basis. If a packet didn't make it to the final destination, it would shortly be retransmitted from the source.
- Black boxes would be used to connect the networks; these would later be called gateways and routers. There would be no information retained by the gateways about the individual flows of packets passing through them, thereby keeping them simple and avoiding complicated adaptation and recovery from various failure modes.
- There would be no global control at the operations level.

Whether an information infrastructure should be based on a universal network design or allow the degree of heterogeneity underlying the Internet technology was the source for heated debates throughout the seventies (and eighties) in the discussions about standards for such infrastructures. In this debate, one alternative was the Internet and its TCP/IP suite, the other the X.25 standard proposed by CCITT and telecom operators (Abbate 1995). X.25 was put forth as the universalistic alternative to TCP/IP.

The telecom operators preferred X.25 as it was design according to the ideology of universalism, having a strong tradition since the very early days of telecommunication. However, such a homogeneous, universalistic solution was also well aligned with their interest in extending their monopoly from telephone communication and into computer communications (ibid.).

However, the Internet community is far from consistent in its thinking related to universalism. The supporters of the Internet technology, including TCP/IP, have always argued for TCP/IP as the universal network/transport level protocol. The rationale behind TCP/IP is an example of basic assumptions about a heterogeneous, pluralistic world with diverging needs, requiring different network technologies (like radio, satellite and telephone lines). However, this description of a heterogeneous, open world is put for as an argument for one specific universal solution, TCP/IP. From the level of TCP/IP and up there is no heterogeneous world any more, only consistent and coherent user needs to be satisfied by universal solutions. This view is maybe strongest and most explicitly expressed by Einar Stefferud when arguing against e-mail gateways (Stefferud 1994). He argues that gateways translating between different e-mail (address and content) formats and protocols should not be "allowed" as such translations cannot in principle be perfect. In such translations there is, at least in principle, possibilities for loss of information.

Stefferud proposes another solution for integrating enclaves of users using separate e-mail systems. He calls this principle *tunnelling*, meaning that an e-mail message generated by one e-mail system, say cc:mail, might be tunnelled trough an enclave of e-mail systems of another kind, say X.400 systems, by enveloping the original e-mail in an e-mail handled by the second e-mail system. However, this technique can only be used to send an e-mail between users of the same kind of e-mail systems. A user connected to a cc:mail system might send a mail to another cc:mail user through, for instance, a X.400 network. However, it will not allow communication between a user of a cc:mail system and a X.400 user. If tunnelling is a universal solution, as Stefferud seems to believe, it presupposes a world built up of separate sub-worlds, between which there is no need for communication. Universalism and closed world thinking have also strong positions in the Internet community.

Science: universal facts and laws

Universalism has its strongest position in science. Universalism in technological development is usually the result of seeing scientific ideals as the ultimate, and accordingly trying to apply scientific methods in technological development.

The traditional view on science is that science is simply the discovery of objective, universal facts, laws and theories about nature (and possibly other worlds, like the social). In this sense, universalism corresponds to a kind of heliocentrism, namely the (implicit) assumption that your own position is priviledged, that you are the origo around which everyone circle. It would carry us well beyond the scope of this book to pursue universalism within science. For our purposes it suffices to observe the many forms and appearances of universalism as well as its far-reaching influence.

Standards are found everywhere, and as such they have been focused. Standards - in a wide sense- are indeed *the* issue addressed in STS. Not specifically technological standards, but rather standards in form of universal scientific facts and theories. These studies also, we believe, have something to tell us about information infrastructure standards.

0.1. constructing scientific facts, theories, technologies, standards

Universality, actor network theorists have argued, is not a transcendent, a priori quality of a body of knowledge or a set of procedures. Rather, it is an acquired quality; it is the effect produced through binding heterogeneous elements together into a tightly coupled, widely extended network. In his elegant study on the creation of universality, Joseph O'Connel discusses the history of electrical units.[2] Laboratory scientists, US war planes and consumers buying new TV's do not simply plug into some pre-given, natural Universal called the Volt. Rather, the Volt is a complex historical construct, whose maintenance has required and still requires legions of technicians, Acts of Congress, a Bureau of Standards, cooling devices, precisely designed portable batteries, and so forth.

Other theoretical traditions within STS likewise question these rhetorics. Social constructivist analyses, for example, also argue that the universality of technology or knowledge is an emergent property: to them, a fact or a technology becomes universal when the relevant social actors defining it share a common definition.

0.2. obtaining universality

- deleting context
- Geof

The maybe most basic finding within STS is the *local* and situated nature of all knowledge - including scientific knowledge. Latour and Woolgar (1986) describes how scientific results are obtained within specific local contexts and how the context is deleted as the results are constructed as universal. Universals in general (theories, facts, technologies) are constructed as the context is deleted, basically by being taken as given. This construction process has its opposite in a deconstruction process when universals are found not to be true. In such cases the universal is deconstructed by re-introducing its context to explain why it is non valid in the context at hand (Latour and Woolgar, 1979).

In spite of the fact the context of origin and the interests of its originators are “deleted” when universals are created, these elements are still embedded in the universals. They are shaped by their history and not just objectively reflecting some reality (in case of scientific facts of theories) or being neutral tools (in case of universal technologies). They embed social and political elements.

In the same way as other universals, infrastructure are standards in fact “local” (Bowker and Star 1994, Timmermans and Berg 1997). They are not pure technical artifacts, but complex heterogeneous actor-networks (Star and Ruhleder 1996, Hanseth and Monteiro 1997). When a classification and coding system like ICD

(International Classification for Diseases)[2] is used, it is embedded into local practices. The meaning of the codes "in use" depends on that practice (Bowker and Star 1994). The ICD classification system developed and maintained by WHO in order to enable a uniform registration of death causes globally (to enable the generation of statistics for research and health care management) reflects its origin in the Western modern world. "Values, opinions, and rhetoric are frozen into codes" (Bowker and Star 1994, p. 187). Common diseases in the third world are less well covered, and the coding system is badly suited for the needs of a third world health care system.

0.3. implementation - making it work

- using a theory
- implementing/using a standard

In parallel with showing how universals are constructed, STS studies have addressed maybe more extensively how they are used, i.e. how they are made to work when applied in spite of the seemingly paradoxical fact that all knowledge is local. This is explained by describing how the construction of universal, the process of universalization, also has its opposite, the process of localization. The meaning of universals in specific situations, and within specific field, is not given. It is rather something that has to be worked out, a problem to be solved, in each situation and context. If we will apply a universal (theory) in a new field, how to do that properly, might often be a rather difficult problem to solve, or in other words - working out the relations between the universal and the local setting is a matter of a challenging design issue. As a universal is used repeatedly within a field (i.e. community of practice), a shared practice is established, within which the meaning and use of the universal is taken as given.

boundary objects

practical accomplishment

Lucy plans

Just as the development of universal is not a neutral activity are social and political issues involved in the use of universals. As their use is not given, "designing" (or

2. ICD has been defined under the authority of WHO, and is used in health care institutions in most of the world. ......

"constructing") the use of universals is a social activity like any others, taking place within a local context where social and political issues are involved.

Marc Berg and Stefan Timmermans argue that studies in the STS field tend to reject the whole notion of universals (Timmermans and Berg 1997, Berg and Timmermans 1998). They disagree, saying that universals exist, but they are always embedded into local networks and infrastructures. Universals exist - but as *local universals*. "The core of universalities lies in the changes built up on local infrastructures." They argue further that there are always multiplicities of universalities. Some of these will be in conflict. Each universal defines primarily an order it is meant to establish. Implicitly it defines at the same time *dis*-order - does not match the standard. When a multiplicity of standards are involved in an area - which is "always" the case - on standard's order will be another's dis-order. Further, Berg and Timmermans show how a standard even contains, builds upon, and presuppose dis-order.

In this paper we will use these theories and concepts to discuss the definition, implementation and use of corporate infrastructure standards. We will do this by first showing how the need for a universal solution - a standard - was constructed, and subsequently the decision to define and implement a corporate standard called Hydro Bridge, and the definition of its content.

The main part of the article is concentrating on the implementation of the standard. The most characteristic aspect of this implementation of the process is the repeated discovery of the incompleteness of standard in spite of all efforts to extend it to solve this very incompleteness problem. This process is a continuous process of enrolling new actors and technological solutions to stabilize the network constituting the standard. This stabilization process never terminates - partly due to the open nature of infrastructures, but may be more important because the standard creates disorder within exactly the domain it is designed and implemented in order to bring into order.

The origin of universalism

Constructing the need for universal standards

It is not at all obvious that solutions should be maximally applicable, so where does the idea stem from? We illustrate this by drawing on the experience with developing a Norwegian health information infrastructure and tracing its initial, international efforts.

The choice of a standardisation model was not given from the outset. The general *Zeitgeist*, however, was that of working out as universal and open standards as possible as explained earlier for lab communication (see chapters 2 and 7). Adopting EDIFACT as the basis for electronic prescriptions seemed inevitable even though alternatives were proposed. These alternatives inscribe quite different interests and delegate completely different roles and competencies to involved actors, especially the EDIFACT mafia.

There were several, alternative standardisation and information infrastructure development strategies, or models, promoted originally. These models are all more or less based on deep-seated convictions about how technology development takes place. They inscribe quite different spheres of authoritative competence and steps to proceed in the design. The range of technically feasible standardisation models was practically unlimited. This implied that deciding on one model was less a question of technical superiority of any one model and more a question of who should be allowed to function as a gatekeeper in defining the problem.

The development of electronic information exchange between health care institutions in Norway started when a private lab, Dr. Fürst's Medisinske Laboratorium in Oslo, developed a system for lab report transmission to general practitioners in 1987. The system was very simple — the development time was only 3 weeks for one person. The interest of Dr. Fürst's laboratory was simply to make profit by attracting new customers. It was based on the assumption that the system would help general practitioners save much time otherwise spent on manual registering lab reports, and that the general practitioners would be find this attractive. Each general practitioner receives on average approximately 20 reports a day, which take quite some time to register manually in their medical record system.

The system proved to be a commercial success and brought them lots of general practitioners as new customers. This implied less profit for the other labs. Within a couple of years, several non-private labs (in hospitals) developed or bought systems with similar functionality in order to be competitive. Although these systems were more or less blue-prints of that of Dr. Fürst's laboratory, there were differences which inscribed extra work for the vendors of electronic medical record systems for the general practitioners. This gave these vendors incentives for working out one, shared solution.

Alongside the growing number of labs adopting systems for exchange of reports, an increasing number of actors saw a wider range of applications of similar technology in other areas. These actors were represented within the health sector as well as among possible vendors of such technology. For all of them it was perceived as important that the technologies should be shared among as many groups as possible in order to reduce costs and enable interconnection of a wide range of institutions.

Telenor (the former Norwegian Telecom) had strong economical interests in promoting extensive use of tele- and data communication based services. As telecommunication technology became more integrated with IT, Telenor searched for candidates for extensive use of new and advanced services. The health sector was selected as the potentially most promising one. After an initial experiment, Telenor launched the project “Telemedicine in Northern Norway” in 1987 which was running until 1993. Although Telenor realised that the services and products developed for a specific sector like health care could never be as general as the telephone, Telenor had a strong economical incentive to make their market as large as possible. This strategy presupposes that the standards are as general as possible in order to cover as many sectors as possible.

Standardisation has always been considered important within the telecommunication sector. Hence, Telenor took it for granted that the new health information infrastructure standards should be like any other telecommunication standard: “open” and developed according to the procedures of formal standardisation bodies. Telenor effectively acted as a standardisation “partisan”. Their perceived neutrality together with the investments in the telemedicine project made Telenor a very influential actor within information infrastructure standardisation in Norway in the 80s.

The historical legacy of telecom

Universalism within telecom has a long-standing history. It was first coined in 1907 by the president of AT & T, Theodor Vail, and amounted to “one policy, one system and universal service” (Mueller 1993, cited in Taylor and Webster 1996, p. 219). The notion of universalism in telecom is not well defined. It started out as a tidiness principle, namely the principle of a unified, non-fragmented service. It has since come to include also issues of coverage and reach. The heritage of universalism in telecom mirrors the deeply felt obligation of early telecom providers to avoid fragmentation and inequity. It thus inscribed clear, political goals. Universalism in telecomm was heavily influenced — arguable even crucially dependent upon — the prevailing monopoly situation (Taylor and Webster 1996, p. 220):

> “The key to this construction of universal service was that it linked political goals, such as universal service, to a particular system of economic

> organisation, a monopoly which sustained itself through revenue-pooling arrangements."

The question, then, is how universalism may, or indeed should, unfold in the current situation with increasing deregulation.

Within a region controlled by a telecom operator, the ideal of equity was pronounced (XXREF soc hist of tele). Telephone service should be as general as possible, everyone should be granted the same opportunities of access and service. Abbate (1995) shows how this biases the telecom world towards centralization as a mean to achieve coherence, consistency and non-redundancy.

In the currently ongoing deregulation of the telecom sector, "universal service" is a key term. This terms means that the deregulation needs to happen in a way guaranteeing universal service, i.e. all services provided to anybody in a country should be provided to everybody (at the same price) all over the country (OECD).

How to create universals

Having sketched the background for universalism as well as some of its expressions, let us explore next how this maps onto practical design efforts. How, then, do designers go about when (implicitly) influenced by universalism?

Making universal health care standards

There is a strong tendency to aim at solutions that, implicitly or explicitly, are to be fairly stable. The drive is to get it right once and for all, to really capture "it":

> "The complete model... will be very big, very complex and expensive to make. In the process of coming there, a collection of smaller, internally consistent sub-models, coordinated via the most common medical objects, and specialized for states and countries, have a value of their own" (MedixINFO).

Object oriented techniques were supposed to provide the tools necessary to develop this all-encompassing information model while all the time keep it consistent and backwards compatible (Harrington 93). The requirements for the model was described as follows:

> "The MEDIX framework requires a medical information model. It must be a conceptual model, describing how actual people and medical systems share and communicate information. It is not a model describing how an automated version of the health care environment distributes and communicates technical representations of medical documents. No computers or communication nodes will be present in the model, only real world objects like patients, physicians, beds, and tests.
>
> Making a common information model for use in MEDIX is a necessary task, involving many people and work-months. On one hand, the model must be precise enough to be used as the basis of operational, communicating health care systems. On the other hand, it must capture the reality as it is perceived by many people" (ibid., p 5).

The information model was intended to be developed and specified according to Coad and Yourdon's "Object-Oriented Analysis" (Coad and Yourdon 1991).

The computational/communication model is an implementation of the information model which is assumed to be done more or less automatically from the specification of the information model. An automated health care IT system is assumed to represent "the underlying health care reality in terms of a computational model of that reality" (Harrington 1990a).

> "The information model serves two purposes. First, it represents the information flow patterns in the health care environment. In this representation, there are no computers or communicating computer processes, only real world entities like patients, physicians and service requests...
>
> Once the information model has been defined and validated, however, it takes on a new mission in which the model must relate to information technology, and in particular to communication standards and profiles. For the model to serve this mission, it must be translated into precise languages which can be used by automated processes to define the messages and trigger events of the computer systems. Therefore a set of translating (or mapping) rules and methods has been developed for translating selected aspects of the real world model into communication profiles" (ibid.)
>
> "By extracting a small subset of the model, and relating it to a single medical scenario, it is possible to "prove" the model's suitability and correct-

> ness with respect to this tiny subset, if a set of medical experts agree that it is a “correct” representation of the real world” (ibid., p. 12).

The MEDIX framework is a strict and straight-forward application of the classical information modeling approach to information systems development. Some weaknesses of this approach are mentioned. However, these weaknesses are implicitly assumed to be irrelevant as there is no attempt to deal with them.

The key tool and strategy for creating universal standards is information modelling. This tool and strategy directly mirrors a naive realist position (Hanseth and Monteiro 1994 (XXSJIS)) within the theory of science, believing that the objective true world is discovered of we use proper methods, and that the world discovered using such a method is a consistent one of manageable complexity. These assumptions is extensively used in the rhetorics of information modelling, in particular when arguing its advantages over alternative positions.

As mentioned, some weaknesses with the approach are mentioned. However, these are no serious weaknesses - either they are not that serious, or they can be solved. The complexity of the models searched for is one such possible problem. However, this is no real problem. It is a problem only to the extent in can be solved, i.e. the only problems seen are problems that really aren’t problems. Or problems are only seen when the solution appears. The complexity problems exist only to the extent that object oriented technique can solve them (Harrington 1993).

Universals in practise

So far our analysis and critique of universalism has been of a conceptual nature. If we bracket this theoretically biased view for a moment, what is the practical experience with standardisation dominated by universalism? Is a the critique of universalism but high-flying, theoretical mumble void of any practical implications?

Evaluation at a quick glance

The massively dominant approach to date has been met with surprisingly few objections. The heritage from telecommunication standardisation and information modelling (see above) is evident in the thinking and actions of the EDIFACT mafia. It was, for instance, simply “obvious” that problem of developing lab messages in Norway should be translated from acquiring practical experience from situations of use in Norway to aligning the specification with perceived European requirements. The EDIFACT mafia had a gatekeeping role which allowed them to define the problem. And their definition of the problem was accepted. Proponents of alternatives (for instance, Profdoc´s bar codes) was incapable to market their solutions to users. The statement from EDIFACT cited in (Graham et al. 1996, p.10, emphasis added) illustrates how problems are down-played and benefits are exaggerated: "It should be understood that the benefits of having a single international standard outweigh the *drawbacks of the occasional compromise*".

The diffusion, in line with (Graham et al. 1996), has been very slow. The non-standardised lab message systems developed and adopted by users in the period 1987 to 1992 are still in use although their further diffusion has stopped. The installations of systems based on standardised lab messages seem to be used as described by the scenarios worked out as part of the standardisation work. Similarly, the EDIFACT messages implemented adhere to the practice inscribed into the actor-network constituting EDIFACT technology. There is no implementations of the standards based on re-interpretations of some of the design assumptions. Dr. Fürst's lab consider implementing a systems providing services beyond what can be offered by a standardised one as being too difficult at the moment. This would require cooperation with other actors, and establishing such an arrangement is too difficult as it is based on an anti-program compared to that inscribed into the standards.

Learning from experience?

CEN's own judgement is that "CEN/TC 251 has so far been a successful Technical Committee." (CEN 1996, p. 3). This is obviously true from a purely political point of view in the sense that it has established itself as the most authoritative standardization committee on the European level within the health care area. Looking at the implementation (diffusion) of the standards defined, the judgement may be different. So far, ten years of standardization work within Medix and CEN has hardly had any effect on information exchange within health care. Maybe the most significant effect has been that the standardization work has made everybody awaiting the ultimate standards rather then implementing simple, useful solutions. Within the health care sector in Norway, the simple solutions for lab report exchange systems diffused very fast around 1990. The standardization efforts seem to have stopped rather than accelerated the development and diffusion of IIs.

The CEN approach is an example of those being criticized for being all to slow and complex, not satisfying user needs. The Medix work started in a period with limited experience of the kind of work it was undertaking. In the CEN case however, more experience is available. This experience does not seem to influence CEN's approach, neither has the discussion about strategies for implementing NII and the Bangemann plan.

CEN considers the development of consistent, non-redundant global standards most important for building IIs. User influence is also considered mandatory. However, it believes user participation in the information modelling work will do the job. It does not consider changing the standards to adapt to future needs to be of any importance. The issue is mentioned but not addressed. Even the brief comments made in Medix about the need for evolution and how object-oriented methods would enable this has disappeared. CEN's view on their own work is in strong contrast to how they look at ongoing local activities implementing IIs for specific needs. This work is considered of greatest danger as it will lead to permanently incompatible IIs. Accordingly, CEN is based on a belief that these smaller and simpler IIs cannot be changed and evolve into larger interconnected networks. How their own standards, and the IIs implementing them, is going to avoid this problem, i.e. the difficulties in being changed to accommodate to future needs, is hard to see.

The view on standards found in the health care standardization communities mentioned above is quite common among those involved in standardization. For instance, within CEC's Telematics Programme a similar method for engineering trans-European telematics applications is developed (Howard, 1995).

The illusion of universalism

Universalism is an illusion, at least in form of universal information infrastructure standards. This fact will be illustrated by the openness of the use and use areas of infrastructures, the unavoidable duplication and the incompleteness of any information infrastructure standard.

Openness: However universal the solutions are intended to be, there are still more integration and interconnection needs that are not addressed. At the moment of writing this, a popular Norwegian IT journal raises criticism against the government for lack of control ensuring necessary compatibility (ComputerWorld Norge 1997). The problem is, it is said, the fact that two projects are developing solutions for overlapping areas without being coordinated. One is developing a system, including a module for digital signatures, for transmission of GPs' invoices. The (digital signature) security system is intended to be used in all areas for communication between social insurance offices and the health sector as well as for internal health sector communication. It is designed according to CEN specifications as far as possible, to be compatible with future European II needs within health care. The other project develops a solution for document exchange with the government, including a system for digital signatures supposed to be the "universal system" for the government sector - which the social insurance offices are parts of.

This example illustrates a dilemma which cannot be solved, neither can one escape from it: Should CEN specify security systems standards not only for the health care sector in Europe, but the whole public sector as well? Or should the development of a security system for the government sector also include the solution for the whole health sector as well - including trans-national information exchange and accordingly being compatible with the needs in all European countries? And what about the other sectors anybody involved communicates with?

Incompleteness: When developing a “universal solution,” irrespective of its completeness it will always be incomplete is the sense that its specification has to be extended and made more specific when it is implemented. For instance, when two partners agree on using a specific EDIFACT message, they must specify exactly how to use it. Although a standard EDIFACT message is specified and so is the use of X.400 as standard for the transport infrastructure, other parts are missing. Such parts may include security systems. If a security system is specified, for instance one requiring a TTP (i.e. a “trusted third party”), the security system has to be adapted to the services supported by the TTP, etc. All these “missing links” introduce seams into the web. This made the implemented solution locally situated and specific - not universal. The universal solution is not universal any more when it is implemented into reality.

Duplication: When implementing a solution, “real” systems are required, not just abstract specifications. Usually the nodes in an II will be based on commercially available implementations of the standards chosen. However, to work together, commercial products have to be adapted to each other. In an EDIFACT based solution requiring security services (encryption, digital signatures, etc.), the so-called EDIFACT converter and the security system must fit together. Usually each converter manufacturer adapt their product to one or a few security systems providing some security functions. This means that when an organization is running two different EDIFACT based services with different security systems, they will often not only have to install two security systems, but two EDIFACT converters as well as the security systems are not integrated with the same EDIFACT converter. This problem is found in the implementation of the EDIFACT solutions for lab reports and GPs’ invoices respectively. The GPs using both kinds of services must install two separate systems, duplicating each other’s functionality completely: Two EDIFACT converters, two security systems, two X.400 application clients and even access to two separate X.400 systems and networks! (Stenvik 1996).

The problems the development of “universal solutions” meets are basically exactly those the believers in universal solutions associate with heterogeneous solutions, and which the idea about the “universal solution” is proposed to solve. In some cases the solutions developed is even worse than “non-standard” ones because the assumed standardized solutions are equally heterogeneous and in addition much more complex.

This kind of inconsistency is also found within CEN concerning installed base issues. Those involved appear to be well aware of the irreversibility phenomena as far as existing local solutions are concerned (De Moor 1993, McClement 1993). The irreversibility of these incompatible solutions is one CEN's most important arguments in favour of universal, consistent and non-redundant standards. However, it is a bit strange that they apparently believe that the phenomenon does not apply to their own technology.

Linking variants

Universalism faces serious and greatly under-estimated problems. In practise, especially for working solutions, the design of consistent non-redundant solutions inherent in universalism has been abandoned. Instead, duplication, redundancy and inconsistency are allowed. The situation is kept at a manageable level by linking networks through gateway mechanisms.

Successful information infrastructure building efforts have not followed the ideology of universalism, but rather an opposite "small is beautiful" approach. This approach has been followed by concentrating on solving more urgent needs and limited problems. The system developed by Fürst and later copied by lots of other labs is an example of this. To my knowledge, there is no larger effort in the health sector explicitly trying to implement larger IIs based on explicitly chosen transition and interconnection strategies and gateway technologies. However, smaller networks are connected based on shorter term perspectives and practical approaches. Fürst, for instance, has connected their system for lab transmission to GPs in Norway to the system used by three UK and US based pharmaceutical companies (receiving copies of lab reports for patients using any of their drugs being tested out). The networks interconnections are based on "dual stack" solutions. Interfacing a new network and message format takes about one man week of work (Fiskerud 1996).

The Fürst experience indicates that building larger IIs by linking together smaller networks are rather plain and simple. The difference in complexity between the Fürst solution and the CEN effort is striking. The specification of the data format representing lab reports used in the Fürst system covers one page. The CEN specification of the EDIFACT message for lab reports covers 500 pages. Its specification work lasted for 4 to 5 years. According to one Norwegian manufacturer of medical record systems for GPs, for two partners that want to start using a standardized solution, ensuring that they are interpreting the CEN standardized message consistently demands just as much work as developing a complete system like Fürst's from scratch.

If universalism is dead, what then?

Based on the experiences mentioned above, two kinds of gateways seems to be particularly relevant. One is gateways linking together different heterogeneous transport infrastructures into a seamless web. The other is "dual stack" solutions for using different message formats when communicating with different partners. This is pursued further in chapter 11.

Experiences so far, indicates that implementing and running dual stack solutions is a viable strategy. If a strategy like the one sketched here is followed, implementing tools for "gateway-building" seams to a task of manageable complexity.

In the chapter that follows, we explore how the fallacy of universalism can be avoided by appropriating the inscriptions of the existing bits and pieces of an infrastructure.

CHAPTER 9

Installed base cultivation

Introduction

We have outlined how the elements of an information infrastructure inscribe future patterns of use (chapters 6 and 7) and how the to date dominating approach inscribes completely unrealistic scenarios of use (chapter 8). This leads us to explore further what more realistic ways to intervene, i.e. design, an information infrastructure. This implies a closer analysis of the way behaviour is inscribed in the already existing elements of an infrastructure — the installed base. The direction of our analysis will be a radically different approach to the “design” of infrastructure, an approach dubbed “cultivation”.

Installed base

The building of large infrastructures takes time. All elements are connected. As time passes, new requirements appear which the infrastructure has to adapt to as explained in chapter 5. The whole infrastructure cannot be change instantly - the new has to be connected to the old. The new version must be designed in a way making the old and the new linked together and “interoperable” in one way or another. In this way the old - the installed base - heavily influence how the new can be designed. Infrastructures develops through extending and improving the *installed base*.

The focus on infrastructure as "installed base" implies that infrastructures are considered as always already existing, they are NEVER developed from scratch. When "designing" a "new" infrastructure, it will always be integrated into and thereby extending others, or it will replace one part of another infrastructure. This has been the case in the building of all transport infrastructures: Every single road - even the first one if it make sense to speak about a such - has been built in this way; when air traffic infrastructures have been built, they have been tightly interwoven with road and railway networks - one needed these other infrastructures to travel between airports and the travels' end points. Or even more strongly - air traffic infrastructures can only be used for one part of a travel, and without infrastructures supporting the rest, isolated air traffic infrastructures would be useless.

The irreversibility of the installed base

Information infrastructures are large actor-networks including: systems architectures, message definitions, individual data elements, standardisation bodies, existing implementations of the technology being included in a standard, users and user organisations, software vendors, text books and specifications. Programs of action are inscribed into every element of such networks. To reach agreement and succeed in the implementation of a standard, its whole actor-network must be aligned.

In the vocabulary of actor-network theory (presented in chapter 6), this insight corresponds to recognising that the huge actor-network of Internet — the immense installed base of routers, users' experience and practice, backbones, hosts, software and specifications — is well-aligned and to a large extent irreversible. To change it, one must change it into another equally well-aligned actor-network. To do this, only one (or very few) components of the actor-network can be changed at a time. This component then has to be aligned with the rest of the actor-network before anything else can be changed. This gives rise to an alternation over time between stability and change for the various components of the information infrastructure (Hanseth, Monteiro and Hatling 1996).

Within the Internet community, the importance of the installed base and its heterogeneous, i.e. socio.technical, character are to some extent acknowledged:

> "I would strongly urge the customer/user community to think about costs, training efforts, and operational impacts of the various proposals and PLEASE contribute those thoughts to the technical process."
>
> *(Crocker 1992)*

> "Key to understanding the notion of transition and coexistence is the idea that any scheme has associated with it a cost-distribution. That is, some parts of the system are going to be affected more than other parts. Sometimes there will be a lot of changes; sometimes a few. Sometimes the changes will be spread out; sometimes they will be concentrated. In order to compare transition schemes, you "must" compare their respective cost-distribution and then balance that against their benefits."
>
> *(Rose 1992b)*

In the next chapter we will look at installed base issues involved in the design of the new version of IP, IPv6.

Information infrastructure as actor

A large information infrastructure is not just hard to change. It might also be a powerful actor influencing its own future life - its extension and size as well as its form.

Within the field institutional economy some scholars have studied standards as a part of a more general phenomena labelled "self-reinforcing mechanisms" (Arthur 1988, 1989, 1990) and "network externalities" (Katz and Shapiro 1986). We will here briefly review these scholars conception of standards and the and their self-reinforcing character as the installed base grow. We will look at the cause of this phenomenon as well as its effects.

Self-reinforcing mechanisms appear when the value of a particular product or technology for individual adopters increases as the number of adopters increase. The term "network externalities" is used to denote the fact that such a phenomenon appears when the value of a product or technology depends also on aspects being external to the product or technology itself.

A standard which 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,

p. 2). The development and diffusion of standards and "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." (ibid., p. 7)

The basic mechanism is that the large installed base attracts complementary production 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, etc. (ibid., p. 27).

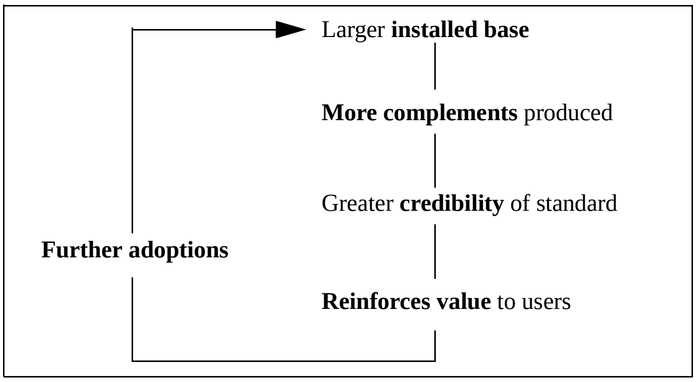

FIGURE 4. **Standards reinforcements mechanism (Grindley 1995).**

Self-reinforcing mechanisms are, according to Arthur (1990), outside the scope of traditional, neo-classical economy, focusing on diminishing return on investment. Typical examples of this phenomenon is found within resource based economics. Assuming that the sources (of for instance hydro power, oil, ..) most easily available are used first, the costs increases as more is consumed (Arthur 1990).

In general, the part of the economy that is knowledge-based are largely subject to increasing returns: large investments in research, incremental production is relatively cheap, increased production means more experience and greater understanding in how to produce additional units even more cheaply, benefits of using them increase (ibid.)

When network externalities are significant, so too are the benefits of having compatible products, and accordingly the establishment of common standards (Katz and Shapiro 1986).

Arthur (1988) identifies four *sources* of self-reinforcing processes: Large set-up or fixed costs; learning effects (improvement through experience); coordination effects (advantages in going along with others); and adaptive expectations. Katz and Shapiro (1985) presents three possible sources of network externalities: Direct physical effect of the number of purchases on the quality of the product (for instance telephone); indirect effects, for instance the more users that buy a particular computer, the more software will be available; and post-purchase service depends on the experience and size of the service network.

The self-reinforcing effects of the installed base corresponds to Thomas Hughes' concept of *momentum*. This concept is one of the important results of Hughes study of electricity in Western societies in the period 1880-1930 (Hughes 1983). This study is certainly the most important and influential within the field of LTS studies.

Hughes describes momentum as very much a self-reinforcing process gaining force as the technical system grows "larger and more complex" (Hughes 1987., 108). Major changes which seriously interfere with the momentum are, according to Hughes, only conceivable in extraordinary instances: "Only a historic event of large proportions could deflect or break the momentum [of the example he refers to], the Great Depression being a case in point" (ibid., 108) or, in a different example, the "oil crises" (ibid., 112). As Hughes describes it, momentum is the result of a larger actor-network including the technology and its users, manufacturers, educational institutions, etc. In particular the establishment of professions, educational institutions and programs are crucial in this respect. We will give a more comprehensive presentation of Hughes study later in this chapter as an example of an evolving infrastructure.

Effects

Arthur (1988) points to some *effects* of self-reinforcing mechanisms:

1. *Path-dependence;* i.e. passed events will have large impacts on future development and in principle irrelevant events may turn out to have tremendous effects.

2. *Lock-in;* i.e. when a technology has been adopted it will be impossible to develop competing technologies. “Once random economic events select a particular path, the choice become locked-in regardless of the advantages of alternatives” (Arthur 1990).
3. *Possible inefficiency;* i.e. the best solution will not necessarily win.

The most widely known example of this phenomenon is the QWERTY layout of typewriter and computer keyboards (David 1986). Other examples documented by economists include the VHS standard for VCRs and water-cooling systems in nuclear power plants as well as FORTRAN. Technological standards in general tend to become locked-in by positive feedback (Farrel and Saloner 1985, 1986, 1992).

In the cases of VCR standards and nuclear power plant cooling Betamax and gas cooling respectively were considered technological superior but became losers in the “competition” due to the alternatives, VHS and water cooling, getting an early “advantage” being reinforced through positive feedback.

IIs and communication technologies are paradigm examples of phenomena where “network externalities” and positive feedback (increasing return on adoption) are crucial, and accordingly technologies easily being “locked-in” and turning irreversible. All factors mentioned above apply. The positive feedback from new adopters (users) is strong. The usefulness is not only dependent on the number of users, in case of e-mail for instance, the usefulness is to a large extent its number of users. The technology become hard to change as successful changes need to be compatible with the installed base. As the number of users grow, reaching agreement about new features as well as coordinating transitions become increasingly difficult. Vendors develop products implementing a standard, new technologies are built on top of it. As the installed base grows, institutions like standardization bodies are established, the interests vested in the technology grow.

An actor-network becomes irreversible when it is practically impossible to change it into another aligned one. At the moment, Internet appears to be approaching a state of irreversibility. Consider the development of a new version of IP mentioned earlier. One reason for the difficulty to develop a new version of IP is the size of the installed base of IP protocols which must be replaced while the network is running. Another major difficulty stems from the inter-connectivity of standards: a large number of other technical components depend on IP. An internal report assesses the situation more precisely as: “Many current IETF standards are affected by [the next version of] IP. At least 27 of the 51 full Internet Standards must be revised (...) along with at least 6 of the 20 Draft Standards and at least 25 of the 130 Proposed Standards.” (RFC 1995, 38).

The irreversibility of II has not only a technical basis. An II turn irreversible as it grows due to numbers of and relations between the actors, organisations and institutions involved. In the case of Internet, this is perhaps most evident in relation to new, commercial services promoted by organisations with different interests and background. The transition to the new version of IP will require coordinated actions from all of these parties. It is a risk that “everybody” will await “the others” making it hard to be an early adopter. As the number of users as well as the types of users grow, reaching agreement on changes becomes more difficult (Steinberg 1995).

Some examples

Installed base hostility: The OSI experience

For a long period there was a fight between OSI and Internet - sometimes called a religious war (Drake 1993). Now the war is obviously over and OSI is the looser. Einar Stefferud (1992, 1994) and Marshall Rose (1992) have claimed that OSI would be a failure due to its “installed base hostility.” They argued that OSI was a failure as the protocols did not pay any attention to existing networks. They were specified in a way causing great difficulties when trying to make them interwork with corresponding Internet services, i.e. link them to any existing installed bases. The religious war can just as well be interpreted as a fight between an installed base and improved design alternatives. As OSI protocols have been discussed, specified and pushed through the committees in ISO, Internet protocols have been implemented and deployed. The installed base won, in spite of tremendous support from numerous governments and CEC specifying and enforcing GOSIPs (Governmental

The Internet and Web

The Internet has been a rapidly growing installed base. The larger the installed base the more rapid growth. The rapid diffusion of WorldWideWeb is an example of how to successfully build new services on an existing installed base.

The growth of the Internet is a contrast to the failed OSI effort. However, there are other examples also being contrasts. Leigh Star and Karen Ruhleder (1994, 1996) documents how a nicely tailored system supporting a world wide research community “lost” for the much less sophisticated (i.e. specialized, tailored to users’ needs) gopher and later Web services. As we see this case, the specialized system lost due to its requirement for an installed base of new technological infrastructure. Important was also the lack of installed base of knowledge and systems support infrastructure.

Strategy

We will now look at some implications the importance and role of the installed base will have for choosing infrastructure design strategies.

Dilemmas

David (1987) points out three strategy dilemmas one usually will face when developing networking technologies and which are caused by installed base issues:

1. *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.
2. *Blind giants.* Governmental agencies (or other institutions making decisions) are likely to have greatest power to influence the future trajectories of network technologies, just when 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.
3. *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.

One strategy David (ibid.) finds worth considering is that of "counter-action" - i.e. to prevent the "policy window" from slamming shut before the policy makers are better able to perceive the shape of their relevant future options. This requires positive action to maintain leverage over the systems rivalry, preventing any of the presently available variants from becoming too deeply entrenched as a standard, and so gathering more information about technological opportunities even at the cost of immediate losses in operations efficiency. In "the race for the installed base" governments could subside only the second-place system.

Solutions - enabling flexibility

According to Arthur (1988), exit from an inferior lock-in in economics depends very much on the source of the self-reinforcing mechanism. It depends on the degree to which the advantages accrued by the inferior "equilibrium" are reversible or transferable to an alternative one. Arthur (1988) claims that when learning effects and specialized fixed costs are the source of reinforcement, advantages are not transferable to an alternative equilibrium. Where coordination effects are the source of lock-in, however, he says that advantages are often transferable. As an illustration he mentions that users of a particular technological standard may agree that an alternative would be superior, provided everybody switched. If the current standard is not embodied in specialized equipment and its advantage-in-use is mainly that of convention (for instance the use of colors in traffic lights), then a negotiated or mandated changeover to a superior collective choice can provide exit into the new equilibrium at negligible cost.

The may be most important remedy to help overcome the negative effects of positive feedback and network externalities, i.e. lock-in and inefficiency, is the construction of gateways and adapters (Katz and Shapiro 1985, David and Bunn 1988). Gateways may connect heterogeneous networks, being built independently or based on different versions of the same standards.

Based on an analysis of the circumstances and consequences of the development of the rotary converter which permitted conversion between alternating and direct current (and vice versa), David and Bunn (1988) argue that ".. in addition to short-run resource saving effects, the evolution of a network technology can be strongly influenced by the availability of a gateway innovation." In this case, this gateway technology enabled the interconnection of previously incompatible networks, and enabled the evolution from a dominating inferior technology into a superior alternative.

On a general level, there are two elements being necessary for developing flexible IIs. First, the standards and IIs themselves must be flexible and easy to adapt to new requirements. Second, strategies for changing the existing II into the new one must be developed together with necessary gateway technologies linking the old and the new. These elements are often interdependent.

The basic principle for providing flexibility is modularization and encapsulation (Parnas 1972). Another important principle is leanness, meaning that any module should be as simple as possible based on the simple fact that it is easier to change something small and simple than something large and complex. Formalization increases complexity, accordingly less formalization means larger flexibility.

Techniques that may be used are to design a new version as an extension of the existing guaranteeing backward compatibility and "dual stacks" meaning that a user that are interested in communicating with users on different networks are connected to both. This is explored further in chapter 11.

From design to cultivation

Accepting the general thrust of our argument, that the elements of an infrastructure inscribe behaviour and that there always already exist such elements, implies a radical rethinking of the very notion of design. Traditional design (implicitly) assumes a degree of freedom that simply does not exist (cf. chapter 8 on design dominated by univeralism). This leads us to explore alternative notions of design.

From design to cultivation

When describing our efforts and strategies for developing technological systems, we usually characterize these efforts as design or engineering. Bo Dahlbom and Lars Erik Janlert (1996) use the notion of construction to denote a more general concept including design as well as engineering. They further use the notion of cultivation to characterize a fundamentally different approach to shaping technology. They characterize the two concepts in the following way:

> [When we] "engage in cultivation, we interfere with, support and control, a natural process." [When] "we are doing construction, [we are] selecting, putting together, and arranging, a number of objects to form a system.....
>
> [Cultivation means that] ...we .. have to rely on a process in the material: the tomatoes themselves must grow, just as the wound itself must heal,..
>
> Construction and cultivation give us two different versions of systems thinking. Construction is a radical belief in our power to, once and for all, shape the world in accordance with our rationally founded goals. Cultivation is a conservative belief in the power of natural systems to withstand our effort at design, either by disarming them or by ruining them by breakdown."
>
> *(ibid. p. 6-7)*

The concept of cultivation turns our focus on the limits of rational, human control. Considering technological systems as organisms with a life of their own implies that we focus on the role of existing *technology itself as an actor* in the development process. This theory focuses on socio-technical networks where objects usually considered social or technological are linked together into networks. The "development organization" as well as the "product" being developed are considered unified socio-technical networks.

Improvisation and drifting

Claudio Ciborra propose the concept of "improvisation" as a concept to understand what is going on in organizations when adopting information technology. He holds that this is a concept much more grounded in individual and organizational process than planned decision making (XXCiborra ICIS96). He describes improvisation as situated performance where thinking and action seem to occur simultaneously and on the spur of the moment. It is purposeful human behavior which seems to be ruled at the same time by chance, intuition competence and outright design.

Wanda Orlikowski use the same concept in her "Improvisational Model for Managing Change" (XXOrlikowski Sloan, ISR). In this model organizational transformation is seen as an ongoing improvisation enacted by organizational actors trying to make sense of and act coherently in the world. The model rests on two major assumptions which differentiate it from traditional models of change: first, changes associated with technology implementations constitute an ongoing process rather than an event with an end point after which the organization can expect to return to a reasonably steady state; and second, various technological and organizational changes made during the ongoing process cannot, by definition, all be anticipated ahead of time.

Through a series of ongoing and situated accomondations, adaptations, and alterations (that draw on previous variations and immediate future ones), sufficient modifications may be enacted over time that the fundamental changes are achieved. There is no deliberate orchestration of change here, no technological inevitability, no dramatic discontinuity, just recurrent and reciprocal variations of practice over time. Each shift in practice creates the conditions for further breakdowns, unanticipated outcomes, and innovations, which in turn are responded to with more variations. And such variations are ongoing; there is no beginning or end point in this change process.

Given these assumptions, the improvisational change model recognizes three different types of change: anticipated, emergent, and opportunity-based. Orlikowski distinguish between anticipated changes -- changes that are planned ahead of time and occur as intended -- and emergent changes -- changes that arise spontaneously out of local innovation and which are not originally anticipated or intended. An example of an anticipated change would be the implementation of electronic mail software which accomplishes its intended aim to facilitate increased and quicker communication among organizational members. An example of an emergent change would be the use of the electronic mail network as an informal grapevine disseminating rumors throughout an organization. This use of e-mail is typically not planned or anticipated when the network is implemented, but often emerges tacitly over time in particular organizational contexts.

Orlikowski further differentiate these two types of changes from opportunity-based changes -- changes that are not anticipated ahead of time but are introduced purposefully and intentionally during the change process in response to an unexpected opportunity, event, or breakdown. For example, as companies gain experience with the World Wide Web, they are finding opportunities to apply and leverage its capabilities in ways that were not anticipated or planned before the introduction of the Web. Both anticipated and opportunity-based changes involve deliberate action, in contrast to emergent changes which arise spontaneously and usually tacitly out of people's practices with the technology over time (Orlikowski, 1996).

These three types of change build on each other over time in an iterative fashion (see Figure 1). While there is no predefined sequence in which the different types of change occur, the deployment of new technology often entails an initial anticipated organizational change associated with the installation of the new hardware/software. Over time, however, use of the new technology will typically involve a series of opportunity-based, emergent, and further anticipated changes, the order of which cannot be determined in advance because the changes interact with each other in response to outcomes, events, and conditions arising through experimentation and use.

Similarly, an improvisational model for managing technological change in organizations is not a predefined program of change charted by management ahead of time. Rather, it recognizes that technological change is an iterative series of different changes, many unpredictable at the start, that evolve out of practical experience with the new technologies. Using such a model to manage change requires a set of processes and mechanisms to recognize the different types of change as they occur and to respond effectively to them. The illustrative case presented below suggests that where an organization is open to the capabilities offered by a new technological platform and willing to embrace an improvisational change model, innovative organizational changes can be achieved.

Emergent change also covers where a series of smaller changes add up to a larger whole being rather different from what one was striving for. Some researcher see organizational structures as well as computerized information systems as emergent rather than designed (Ngwenjama 1997). Defined in this way, the concept of emergent change comes close to what Claudio Ciborra (1996a, 1996b, 1997) calls "drifting". When studying groupware, he found that the technology tends to drift when put to use. By drifting he means a slight or significant shift of the role and function in concrete situations of usage, that the technology is called to play, compared to the predefined and assigned objectives and requirements (irrespective of who plans or defines them, users, sponsors, specialists, vendors or consultants). The drifting phenomenon also captures the sequence of ad hoc adjustments. Drifting can be looked at as the outcome of two intertwined processes. One is given by the openness of the technology, its placticity in response to the re-inventions carried out by the users and specialists, who gradually learn to discover and exploit features and potentials of groupware. On the other hand, there is the sheer unfolding of the actors' "being-in-the-workflow" and the continuous stream of bricolage and improvisations that "color" the entire system lifecycle.

Drifting seams to lie outside the scope of control of the various actors; it consists of small and big surprises, discoveries and blockages, opportunistic turns and vicious circles.

Marc Berg (1997) describes drifting in terms of actor network theory. He sees drifting of actor-networks as a ubiqitous phenomena. A network drifts by being unconsciously changed as the effect of a conscious change of another one network whose elements are also parts of others.

Who is controlling whom? technology or humans?

In most discussion about conceptions of and approaches to technological development and change, a key issue is to what extent and how humans can control this process. The concepts of design and construction implicitly assumes that the technology is under complete human control. Improvisation also sees the humans as being in control although less so as the design process cannot be planned. When Orlikowski sees improvisation as "situation change," one might say that "situations" influences the design. However, she does not discuss (or describe) the "nature" of "situations" or whatever might determine what "situations" we will meet in a design or change activity. She notes however that "more research is needed to investigate how the nature of the technology used influences the change process and shapes the possibilities for ongoing organizational change" (Orlikowski xx, p. yy), implying that she assumes that the technology may be an influential actor.

The notion of "drifting" implies that there is no conscious human control over the change process at all. Ciborra does not say anything either about whether there is any kind of "control" or what happens is completely random. However, he is using the notion of technologies being "out of control," which is often used as a synonymous with “autonomous technology,” i.e. technological determinism (Winner 1977). Technological determinism is the opposite extreme of engineers conception of deign and engineering, assuming that it is not humans that design the technology, but rather the technology and its inner logic that determine, i.e. “design,” its own use, implying in the end the whole society - including its own future development. In this extreme position the technology is the only actor.

These two extreme positions is an example of the old dichotomy and discussions within the social sciences between agency and structure. The notions of cultivation is here seen as a middle position where technology is considered shaped by humans although the humans are not in complete control. The technology is also changed by “others,” including the technology itself.

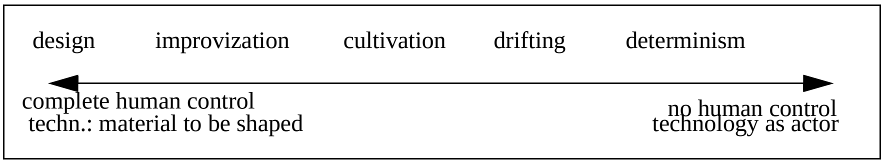

FIGURE 5. Concepts for technological development

Cultivating the installed base

The middle ground

Acknowledging the importance of the installed base implies that traditional notions of design have to be rejected. However, denying humans any role at all is equally, at least, wrong. Cultivation as a middle position captures quite nicely the role of both humans and technology. It is the concept providing us the best basis for developing strategies for infrastructure development. The installed base is a powerful actor. Its future cannot be consciously designed, but designers do have influence - they might cultivate it.

The installed base acts in two ways. It may be considered an actor involved in each single II development activity, but perhaps more important, it plays a crucial role as mediator and coordinator between the independent non-technological actors and development activities.

If humans strive for control, i.e. making our world appropriate for engineering tasks, strategies for cultivating IIs may be considered strategies for fighting against the power of the installed base.

Technological determinism is an extreme position where the installed base is the only actor having power. An actor-network perspective where II development is seen as installed base cultivation is a middle position between technological determinism and social reductionism.

Cultivation might be related to the discussion about the limits and relationships between design and maintenance, and someone might be tempted to see cultivation for just another term for maintenance. This is not the case. First, maintenance is related to design in the way that what has come into being through design is maintained for a while, i.e. until being replaced by a new designed system. Cultivation replaces design in the way that there is nothing being designed which later on will be maintained. Technology comes about and is changed by being cultivated. However, it might still be useful to talk about maintenance as minor changes of minor parts of a larger system or infrastructure.

Transition strategy as cultivation

In terms of actor-network theory, a transition strategy for an information infrastructure corresponds to a situation where one well-aligned actor-network is modified into another well-aligned actor-network. Only one (or a few) of the nodes of the actor-network is modified at a time. At each step in a sequence of modifications, the actor-network is well-aligned. In the more conventional vocabulary of the product world, a transition strategy corresponds to “backwards compatibility” (Grindley 1995). Backwards compatibility denotes the case when a new version of a product — an application, a protocol, a module, a piece of hardware — functions also in conjunction with older versions of associated products. Intel’s micro processor chips, the 80x86 processor family, are examples of backward compatible products. An Intel 80486 processor may run any program that a 80386 processor may run, and a 80386 processor may run any program a 80286 may run. Micorsoft’s Words application is another example of a backwards compatible product: the newer versions of Word may read files produced by older versions of Word (but not the other way around, not forward compatibility).

To appreciate what changing an information infrastructure is all about, it is necessary to identify underlying assumptions, assumptions which go back to the essentially open character of an information infrastructure outlined earlier in chapter 5. It is illuminating to do so by contrasting it with the way changes in another infrastructure technology are made, namely the telephone system. Despite a number of similarities between a telephone infrastructure and an information infrastructure, the issue of how changes are made underline vital and easily forgotten differences.

During the early 90s the Norwegian Telecom implemented a dramatic change in their telephone infrastructure: all telephones were assigned a new 8 digit number instead of the old 6 digit one. The unique digit sequence of a telephone is a core element of any telephone infrastructure. It is the key to identify the geographical and logical location of the phone. Both routing of traffic and billing rely on the unique number sequence. Hence changing from 6 to 8 digits was a major change. The way it was implemented, however, was simple. At a given time and date the old numbers ceased to function and the new ones came into effect. The crucial aspect of this example of a changing infrastructure technology is the strong assumption about the existence of central authority (the Norwegian Telecom) with absolute powers. None were allowed to hang on to the old numbers longer than the others, *absolutely* everyone had to move in concert. This situation might prevail in the world of telecommunication but it is fundamentally antagonistic to the kind of openness hardwired into the notion of an information infrastructure. Making changes in the style of telecommunication is simply not an option for an information infrastructure. There is simply no way to accomplish abrupt changes to the whole information infrastructure requiring any kind of overall coordination (for instance, so-called flag-days) because it is “too large for any kind of controlled rollout to be successful” (Hinden 1996, p. 63). It is accordingly vital to explore alternatives. Transition strategies is one of the most important of these alternatives. To develop a firmer understanding of exactly how large changes can be made, when they are appropriate and where and in which sequence they are to be implemented, is of vital concern when establishing a National Information Infrastructure (IITA 1995).

Strategies for scaling, which necessarily include changes more generally, to information infrastructures are recognized as central to the NII initiative. In a report by the Information Infrastructure Technology and Application working group, the highest level NII technical committee, it is pointed out:

> We don’t know how to approach scaling as a research question, other than to build upon experience with the Internet. However, *attention to scaling as a research theme is essential* and may help in further clarifying infrastructure needs and priorities (...). It is clear that limited deployment of prototype systems will not suffice (...) (IITA 1995, emphasis added)

Contributing to the problems of making changes to an information infrastructure is the fact that it is not a self-contained artefact. It is a huge, tightly interconnected yet geographically dispersed collection of both technical and non-technical elements. Because the different elements of an information infrastructure is so tightly interconnected, it becomes increasingly difficult to make changes when it expands. The inertia of the installed base increases as the information infrastructure scales as is the case with Internet: “The fact that the Internet is doubling in size every 11 months means that the cost of transition (...) (in terms of equipment and manpower) is also increasing.” (IPDECIDE 1993). But changes, also significant ones, are called for.

The scaling of an information infrastructure is accordingly caught in a dilemma. It is a process where the pressure for making changes which ensure the scaling have to be pragmatically negotiated against the conservative forces of the economical, technical and organisational investments in the existing information infrastructure, the installed base. A feasible way to deal with this is for the information infrastructure to evolve in a small-step, near-continuous fashion respecting the inertia of the installed base (Grindley 1995; Hanseth, Monteiro and Hatling 1996; Inhumane and Star 1996; Star and Ruhleder 1996). Between each of these evolutionary steps there has to be a *transition strategy,* a plan which outlines how to evolve from one stage to another.

In the next chapter, we empirically describe an effort to design according to the principles of cultivating the installed base, namely the revision of the IP protocol in Internet. That chapter fleshes out some of the programatically stated principles outlined in this chapter.

CHAPTER 10

Changing infrastructures: The case of IPv6

Introduction

As expectations and patterns of use of an information infrastructure tend to evolve during its life span (see chapter 5), changes are called for. Both minor and major changes are required. This creates a dilemma. The pressure for making changes have to be pragmatically negotiated against the conservative forces of the economical, technical and organisational investments in the existing information infrastructure, the installed base (see chapter 9). A feasible way to deal with this is for the information infrastructure to evolve in a small-step, near-continuous fashion respecting the inertia of the installed base (Grindley 1995; Hanseth, Monteiro and Hatling 1996; Neumann and Star 1996; Star and Ruhleder 1996). Between each of these evolutionary steps there has to be a *transition strategy*, a plan which outlines how to evolve from one stage to another. The controversies over a transition strategy are negotiations about *how* big changes can — or have to — be made, *where* to make them, *when* and in which *sequence* to deploy them.

A transition strategy is a conservative strategy. Rather than trying anything adventurous, it plays it safe. Only modest changes are possible within a transition strategy. A transition strategy is an instance of a cultivation strategy for information infrastructures. In the next chapter we explore other, non-cultivation based approaches to establishing information infrastructures. These approaches facilitate more radical changes to the infrastructure than cultivation based one like the transition strategy described in this chapter.

The IP protocol

The revision of the internet protocol (IP) in the Internet was a direct response to the problems of scaling Internet: “Growth is the basic issue that created the need for a next-generation IP” (Hinden 1996, p. 62). The IP protocol forms the core of the Internet in the sense that most services, including WorldWideWeb, e-mail, ftp, telnet, archive and WAIS, build upon and presuppose IP.

It is fair to say that it has never been more difficult to make changes in Internet than the revision of the IP. This is because the dilemma outlined above has never been more pressing. The explosive growth of Internet is generating a tremendous pressure for making changes, changes which are so fundamental that they need to be made at the core, that is, in IP. At the same time, these changes are likely to have repercussions on an Internet which have never been as huge, never exhibited a stronger inertia of the installed base. Revising IP is the most difficult and involved change ever made to the Internet during its near 30 years of existence. Accordingly, it provides a critical case when studying the problems of changing large information infrastructures.

Our intention is to spell out some of the pragmatics played out within a transition strategy. Adopting a transition strategy is not straightforward. There are a number of socio-technical negotiations that need to be settled. To learn about how to adopt transition strategies accordingly extends well beyond merely stating a conservative attitude. It is necessary to inquire closer into what this amounts to. Changes are always relative. When close, pressing your nose against them, all changes seem big. What, or indeed, whom, is to tell “small” (and safe) changes from “big” (and daring) ones?

Late 80s - July 1992

Framing the problem

There was during the late 80s a growing concern that the success of Internet, its accelerating adoption, diffusion and development, was generating a problem (RFC 1995, p. 4). No one had ever anticipated the growth rate of Internet. The design of Internet was not capable of handling this kind of growth for very long.

Internet is designed so that every node (for instance, a server, PC, printer or router) has an unique address. The core of the problem was considered to be that IPv4 has a 32 bit, fixed length address. Even though 32 bits might theoretically produce 2**32 different identifiers which is a very significant number, the actual number of available identifiers is dramatically lower. This is because the address space is hierarchically structured: users, organisations or geographical regions wanting to hook onto the Internet are assigned a set of unique identifiers (a subnetwork) of predetermined size. There are only three available sizes to choose from, so-called class A, B or C networks. The problem, then, is that class B networks are too popular. For a large group of users, class C is too small. Even though a few times class C would suffice, they are assigned the next size, class B which is 256 times larger than class C.

In this way, the problem of fixed length IPv4 addresses gradually got reformed into the problem of exhausting class B networks. At the August 1990 IETF meeting it was projected that class B space would be exhausted by 1994, that is, fairly soon (ibid., p. 4). This scenario produced a profound sense of urgency. Something had to done quickly. The easy solution of simply assigning several class C network to users requiring somewhat more than class C size but much less than class B was immediately recognised to cause another, equally troublesome, problem. As the backbone routers in Internet, the nodes which decide which node to forward traffic to next, need to keep tables of the subnets, this explosion of the number of class C networks would dramatically increase the size of the routing tables, tables which already was growing disturbingly fast (ibid.). Even without this explosion of class C networks, the size of routing tables was causing severe problems as they grew 50% quicker than hardware advances in memory technology.

During the early 1990s, there was a growing awareness regarding the problems associated with the continued growth of the Internet. It was also recognised that this was not an isolated problem but rather involved issues including assignment policies for networks, routing algorithms and addressing schemes. There was accordingly a fairly clear conception that there *was* a problem complex, but with a poor sense of *how* the different problems related to each other, not to mention their relative importance or urgency. In response, the IETF in November 1991 formed a working group called "Routing and addressing (ROAD)" to inquire closer into these matters.

Appropriating the problem

The ROAD group had by November 1992 identified two of the problems (class B exhaustion, routing table explosion) as the most pressing and IP address exhaustion as less urgent:

> Therefore, we will consider interim measures to deal with Class B address exhaustion and routing table explosion (together), and to deal with IP address exhaustion (separately).
>
> *(RFC 1992, p. 10)*

The two most pressing problems required quick action. But the ROAD group recognised that for swift action to be feasible, changes had to be limited as the total installed base cannot change quickly. This exemplifies a, if not *the*, core dilemma when extending infrastructure technologies. There is pressure for changes — some immediate, others more long-term, some well understood, others less so — which need to be pragmatically balanced against the conservative influence of the inertia of the installed base. This dilemma is intrinsic to the development of infrastructure technology and is accordingly impossible to resolve once and for all. On the hand, one wants to explore a number of different approaches to make sure the potential problems are encountered, but on the other hand one need at one stage to settle for a solution in order to make further progress. It makes more sense to study specific instances of the dilemma and see how it is pragmatically negotiated in every case. A necessary prerequisite for this kind of judgement is a deep appreciation and understanding for exactly how the inertia of the installed base operates.

In the discussions around IPng, the Internet community exhibited a rich understanding of the inertia of the installed base. It was clearly stated that the installed base was not only technical but included "systems, software, training, etc." (Crocker 1992) and that:

> The large and growing installed base of IP systems comprises people, as well as software and machines. The proposal should describe changes in understanding and procedures that are used by the people involved in internetworking. This should include new and/or changes in concepts, terminology, and organization.
>
> *(RFC 1992, p. 19)*

Furthermore, the need to order the required changes in a sequence was repeatedly stated. To be realistic, only small changes can be employed quickly. More substantial ones need to be sugared through a gradual transition.

> The [currently unknown] long-term solution will require replacement and/or extension of the Internet layer. This will be a significant trauma for vendors, operators, and for users. Therefore, it is particularly important that we either minimize the trauma involved in deploying the short-and mid-term solutions, or we need to assure that the short- and mid-term solutions will provide a smooth transition path for the long-term solutions.
>
> *(RFC 1992, p. 11)*

So much for the problem in general. How does this unfold in specific instances? Is it always clear-cut what a “small” as opposed to “large” change is, or what a “short-term” rather than “mid-” or “long-term” solution is? The controversy over CIDR and C# illustrates the problem.

CIDR vs. C#

Instead of rigid network sizes (such as class A, B and C), the ROAD working group proposed employing CIDR (“Class-less Inter-Domain Routing”). CIDR supports variable-sized networks (Eidnes 1994). It was argued to solve many of the problems and that the disruptions to the installed base were known:

> CIDR solves the routing table explosion problem (for the current IP addressing scheme), makes the Class B exhaustion problem less important, and buys time for the crucial address exhaustion problem.
>
> (...) CIDR will require policy changes, protocol specification changes, implementation, and deployment of new router software, but it does not call for changes to host software.
>
> *(RFC 1992, p. 12)*

At this stage, the CIDR solution to the most pressing problems was not well known as Fuller’s(1992) question to the big-internet mailing list illustrates “but what is ‘CIDR´?”. Nor was it unanimous (Chiappa 1992).

Furthermore, alternatives to CIDR existed that had several proponents. One was C# which supported a different kind of variable-sized networks. The trust of the argument for C#, perfectly in line with the fidelity of the installed base, was that it required less changes:

> I feel strongly that we should be doing C# right now. It's not new, and it's not great, but its very easy - there's nothing involved that takes any research, any developments, or any agreements not made already - just say "go" and the developers can start getting this into the production systems, and out into the field. I don't think that CIDR can be done quite that quickly.
>
> *(Elz 1992)*

The discussions surrounding the different short-term solutions for the IP related problems shows broad consensus for paying respect to the installed base. The CIDR vs. C# debate amounts to a judgement about exactly how much changes to the installed base is feasible within a certain time-frame. This judgement varied, producing disagreement and personal frustration. At the same time, the closing down of the controversy and deciding on CIDR illustrates the widespread belief that the need to move on overrides "smaller" disagreements:

> I do feel strongly that it is far more important that we decide on one, and *DO IT*, than continue to debate the merits for an extended period. Lead-times are long, even for the simplest fix, and needs are becoming pressing. So, I want to see us *quickly* decide (agreement is probably too much to ask for :-) on *one* of the three options and *get on with it*!
>
> (...) I will say that I am extremely, deeply, personally, upset with the process that encouraged the creation of the C# effort, then stalled it for months while the Road group educated themselves, leaving the C# workers in the dark, etc., etc.
>
> *(Chiappa 1992)*

The immediate steps including deployment of CIDR was to buy some time badly needed to address the big problem of IP address exhaustion. How to solve the problem was a lot less clear and the consequences were expected to be a lot bigger and cause "significant trauma for vendors, operators, and for users" (RFC 1992, p. 11).

The big heat

At this stage in late 1992, there already had been proposed four solutions to the problem. One solution, called CLNP, was acknowledged to have a certain amount of support but was not accepted (RFC 1992, p. 13). Unable to vouch for any one, specific solution, the IESG only outlined a process of exploration which, hopefully, would lead to a solution. Central to this decision was a judgment about exactly how urgent it was to find a solution. As will become clear further below, this was a highly controversial issue. The IESG position was that there still was some time:

> [T]he IESG felt that if a decision had to be made *immediately*, then "Simple CLNP" might be their choice. However, they would feel much more comfortable if more detailed information was part of the decision.
>
> The IESG felt there needed to be an open and thorough evaluation of any proposed new routing and addressing architecture. The Internet community must have a thorough understanding of the impact of changing from the current IP architecture to a new one. The community needs to be confident that we all understand which approach has the most benefits for long-term internet growth and evolution, and the least impact on the current Internet.
>
> *(RFC 1992, p. 14)*

In parallel with the work of the ROAD group, and apparently poorly aligned with it, the IAB proposed its own plan for the next generation IP (IAB 1992). It was dubbed version 7, written IPv7. This plan of July 1992 opposed the recommendations of the ROAD group and IESG regarding the long-term problem of exhausting IPv4 address space. It produced an unprecedented heated debate during the summer of 1992. The debate focused both on the contents of IAB's solution and decision process producing the plan.

The crucial element of the IAB plan for IPv7 was the endorsement of one of the four available solutions, namely CLNP. The thrust of the argument was appealing to the ideals of Internet design: CLNP existed and people had some experience from it, so why not build upon it? Again, the controversy is not about abstract principles — they are unanimously accepted — but about how to apply the principles to a difficult situation. Hence, the IAB (1992, p. 14) argues that:

> Delaying by a few more months in order to gather more information would be very unlikely to help us make a decision, and would encourage people to spend their time crafting arguments for why CLNP is or is not a better solution than some alternative, rather than working on the detailed specification of how CLNP can be used as the basis for IPv7 (...).

The IAB plan for IPv7 thus makes a different judgement about the available time for the Internet community to search for alternatives than the IESG IPng plan (RFC 1992).

The decisive measures taken by the IAB, settling for a solution rather than keep quarrelling, was praised by a number of people (Braun 1992; Rekhter and Knopper 1992), particularly those close to the commercial interests of Internet. This support for swift action rather than smooth talk is mixed with a discontent for letting the faith of the Internet be left to designers with little or no interest or insight into "reality". A particularly crisp formulation of this position was submitted to the big-internet mailing list shortly after the IAB's decision (Rekhter and Knopper 1992):

> We would like to express our strong support for the decision made by the IAB with respect to adopting CLNP as the basis for V7 of the Internet Protocol.
>
> It is high time to acknowledge that the Internet involves significant investment from the computer industry (both within the US and abroad), and provides production services to an extremely large and diverse population of users. Such and environment dictates that decisions about critical aspects of the Internet should lean towards conservatism, and should clearly steer away from proposals whose success is predicated on some future research.
>
> While other than CLNP proposals may on the surface sound tempting, the Internet community should not close its eyes to plain reality -- namely that at the present moment these proposals are nothing more than just proposals; with no implementations, no experience, and in few cases strong dependencies on future research and funding. Resting the Internet future on such foundation creates and unjustifiable risk for the whole Internet community.
>
> The decision made by the IAB clearly demonstrated that the IAB was able to go beyond parochial arguments (TCP/IP vs. CLNP), and make its judgements based on practical and pragmatic considerations.
>
> Yakov Rekhter (IBM Corporation)
>
> Mark Knopper (Merit Network)

One of the founding fathers of the Internet, Vint Cerf (1992), agreed initially with the IAB that in this case one should organise the efforts rather than fragment them:

> The CLNP specification is proposed as the starting point for the IPv7 both to lend concreteness to the ensuing discussion (I hope this does NOT result in concrete brickbats being hurled through MIME mail....!!) and to take advantage of whatever has already been learned by use of this particular packet format.

But the majority of the Internet was appalled. In the heated debate on the big-internet mailing list, a number of people spoke about "shocked disbelief", "a disastrous idea", "shocked", "dismayed", "strongly disagree" and "irresponsible". The general feeling was clear. The frustration with the decision was obviously very much influenced by the oblique way the IAB had reached its decision thus breaching deep-seated concerns for participating, quasi-democratic decision processes in the Internet.

Bracketing the frustration about the decision process itself, the controversies circled around different views and interpretations of praised design principles.[1] In other words, even though it can be said to be near full consensus among the Internet community regarding concerns about continuity, installed base, transition etc. (see above), the application to specific context is regularly contested. The debate over IAB's IPv7 illustrates this in a striking way.

Abstract design principles meets the real world

The main reason, IAB argued, why it favoured CLNP was that it was necessary for the Internet to find a solution very soon (IAB 1992, p. 14). CLNP is a protocol which "is already specified, and several implementations exist" so it "will avoid design of a new protocol from scratch, a process that would consume valuable time and delay testing and deployment." (ibid., p. 10).

The concern for practical experience is deep and the CLNP solution of the IAB appealed to this. Furthermore, it paves the road for interoperability, another key principle in Internet. Interoperability is recognised to be the end-result of a process of stabilisation:

> I think that relying on highly independent and distributed development and support groups (i.e., a competitive product environment) means that we need a production, multi-vendor environment operating for awhile, before interoperability can be highly stable. It simply takes time for the engineering, operations and support infrastructure to develop a common understanding of a technology.
>
> *(Crocker 1992)*

While acknowledging this design principle, the IAB (1992b) in its Kobe declaration of June 1992 explained its IPv7 decision and argued that for IP an exception had to be made:

> [W]e believe that the normal IETF process of "let a thousand (proposals) bloom", in which the "right choice" emerges gradually and naturally from a dialectic of deployment and experimentation, would in this case expose

1. Alvestrand (1996) suggests that had it not been for the clumsy way IAB announced its decision, many more would probably have gone along with the CLNP solution.

> the community to too great a risk that the Internet will drown in its own explosive success before the process had run its course.

The principal difference was the pragmatic judgement of the amount of time and resources available to work out a revised IP protocol. The IESG's judgement is a head-on disagreement with the IAB's judgment. In addition, more indirect strategies for challenging the IAB were employed. One important line of argument aimed at questioning the experience with CLNP: did it really represent a sufficiently rich source of experience?

> There does exist some pieces of an CLNP infrastructure, but not only is it much smaller than the IP infrastructure (by several orders of magnitude), but important pieces of that infrastructure are not deployed. For example the CLNP routing protocols IS-IS and IDRP are not widely deployed. ISIS (Intra-Domain routing protocol) is starting to become available from vendors, but IDRP (the ISO inter-domain routing protocol) is just coming out of ANSI. As far as I know there aren't any implementations yet.
>
> *(Tsuchiya 1992)*

And more specifically, whether the amount and types of experience was enough to ensure interoperability:

> While there certainly are some implementations and some people using [CLNP], I have no feel for the scale of the usage or -- more importantly -- the amount of multi-vendor interoperability that is part of production-level usage. Since we have recently been hearing repeated reference to the reliance upon and the benefits of CLNP's installed base, I'd like to hear much more concrete information about the nature of the system-level shakeout that it has _already_ received. Discussion about deployment history, network configuration and operation experience, and assorted user-level items would also seem appropriate to flesh out the assertion that CLNP has a stable installed base upon which the Internet can rely.
>
> *(Crocker 1992).*

Interoperability resulting from experience in stable environments presupposes a variety of vendors. CLNP was associated with one specific vendor, DEC, as succinctly coined by Crowcroft (1992): "IPv7 = DECNET Phase 5?" (DECNET is DEC's proprietary communication protocols). Hence, the substance of the experience with CLNP experience was undermined as Crocker (1992) illustrates:

> So, when we start looking at making changes to the Internet, I hope we constantly ask about the _real_ experience that is already widely available and the _real_ effort it will take to make each and every piece of every

> change we require. (...) [R]eferences to the stability of CLNP leave me somewhat confused.

Gaining experience from keeping certain parts stable is a design principle (see above). But some started challenging the very notion of stability. They started questioning exactly what it took for some part to be considered “stable”. An important and relevant instance of this dispute was IPv4. Seemingly, IPv4 has been stable for a number of years as the protocol was passed as an Internet Standard in 1981 without subsequent changes. But even if the isolated protocol itself has been unchanged for 15 years, have there not been a number of changes in associated and tightly coupled elements? Is it, then, reasonable to maintain that IPv4 has been stable?

> How long do we think IP has been stable? It turns out that one can give honestly different answers. The base spec hasn’t changed in a very long time. On the other hand, people got different implementations of some of the options and it was not until relatively recently that things stabilized. (TCP Urgent Pointer handling was another prize. I think we got stable, interoperable implementations universally somewhere around 1988 or 89.)
>
> *(Crocker 1992)*

> I still don’t see how you can say things have been stable that long. There are still algorithms and systems that don’t do variable length subnets. When were variable length subnets finally decided on? Are they in the previous router requirements? (...). So things are STILL unstable.
>
> *(Tsuchiya 1992)*

This is an important argument. It will be addressed also later. In effect, it states that the IP protocol cannot be considered an isolated artefact. It is but one element of a tightly intertwined collection of artefacts. It is this collection of artefacts — this infrastructure — which is to be changed. A shift of focus from the artefact to infrastructure has far-reaching repercussions on what design is all about.

A highly contested issue was exactly which problems CLNP allegedly solved and whether these were in fact the right ones. A well-known figure in the Internet (and OSI) community, Marshall Rose, was among the ones voicing concern that it “is less clear that IPv7 will be able to achieve route-aggregation without significant administrative overhead and/or total deployment.” (Rose 1992a).

After the storm of protests against IAB, combining objections against CLNP with IAB’s decision process, one of Internet’s grand old men, Vint Cerf, reversed the IAB decision at the IETF in July 1992:

> Vint Cerf Monday morning basically retracted the IAB position. They are now supporting the IESG position, and he said that the IAB has learned not to try and enforce stuff from above. (...) Apparently Vint did a strip tease until he took off his shirt to reveal an “IP over everything” T-shirt underneath.
>
> *(Medin 1992)*

The overall result of the hot summer of 1992 was that a plan to explore and evaluate proposals was worked out (RFC 1992). By this time it was clear that “forcing premature closure of a healthy debate, in the name of ‘getting things done’, is *exactly* the mistake the IAB made.” (Chiappa 1992).

July 1992 - July 1994

Let the thousand blossoms bloom, or: negotiating the available time

The situation by July 1992 was this. The IESG recommendation (RFC 1992) of June 1992 calling for proposals drowned in the subsequent controversy over IAB’s IPv7 plan. As the dramatic July 1992 IETF meeting led by Vint Cerf decided to reject the IAB plan, the IESG plan (RFC 1992) was accepted and so a call for proposals for IPng was made at the meeting itself.

The problem now was to organise the effort. Central to this was, again, the issue of time: how urgent were the changes, how many different approaches should be pursued, at which stage should one move towards a closing?

The plan by IESG formulated in June 1992 and revised a month later at the IETF meeting was shaped according to a definite sense of urgency. But it was far from panic. IESG declined to accept the problem as one merely of timing. So even though “[a]t first the question seemed to be one of timing” (RFC 1992, p. 14), the IESG was calm enough to hold that “additional information and criteria were needed to choose between approaches” (ibid., p. 14). Still, the suggested timetables and milestones clearly mirror a sense of urgency. The plan outlines phases of exploring alternatives, elaborating requirements for IPng and a pluralistic decision process — all to be completed within 5 months, by December 1992 (ibid., p. 15). As it turned out, this timetable was to underestimate the effort by a factor of more than by four. It eventually took more than two years to reach the milestone the IESG originally had scheduled for late 1992.

The IESG feared fragmenting the effort too much by spending an excessive amount of time exploring many different proposals. This argument, as illustrated above, was it that led Vint Cerf to initially go along with the IAB IPv7 plan which focused on CLNP. At this stage in July 1992, four proposals existed (called "CNAT", "IP Encaps", "Nimrod" and "Simple CLNP", see (RFC 1995, p. 11)). This was, according to the IESG, more than sufficient as "in fact, our biggest problem is having too many possible solutions rather than too few" (RFC 1992, p. 2).

Following the call for proposals in July, three additional proposals were submitted during the autumn of 1992, namely "The P Internet Protocol (PIP)", "The simple Internet protocol (SIP)" and "TP/IX" (RFC 1995, p. 11). So by the time the IESG had planned to close down on a single solution, the Internet community was facing a wider variety of proposals than ever. Seven proposed solutions existed by December 1992.

Preparing selection criteria

In parallel with, and fuelled by, the submission of proposals, there were efforts and discussions about the criteria for selecting proposals. As it was evident that there would be several to chose from, there was a natural need to identify a set of criteria which, ideally, would function as a vehicle for making a reasonable and open decision.

The process of working out these criteria evolved in conjunction with, rather than prior to, the elaboration of the solutions themselves. From the early sketch in 1992, the set of criteria did not stabilise into its final form as a RFC until the IPng decision was already made in July 1994 (RFC 1994c). It accordingly makes better sense to view the process of defining a set of selection criteria as an expression of the gradual understanding and articulation of the challenges of an evolving infrastructure technology like the Internet.

Neither working on the proposals themselves nor settling for selection criteria was straightforward. The efforts spanned more than two years involving a significant number of people. The work and discussions took place in a variety of forms and arenas including IETF meetings and BOFs, several e-mail lists, working groups and teleconferencing. In tandem with the escalating debate and discussion, the institutional organisation of the efforts was changed. This underscores an important but neglected aspect of developing infrastructure technology, namely that there has to be a significant flexibility in the institutional framework not only (the more well-known challenge of) flexibility in the technology. It would carry us well beyond the scope of this paper to pursue this issue in any detail, but let me indicate a few aspects. The Internet establishes and dismantles working groups dynamically. To establish a working group, the group only has to have its charter mandated by the IETF. In relation to IPng, several working groups were established (including ALE, ROAD, SIPP, TUBA, TACIT and NGTRANS, see ftp://Hsdndev.harvard.edu/pub/ipng/archive/). As the explorative process unfolds during 1993, there is a sense of an escalating rather than diminishing degree of clarity:

> The [IPDECIDE] BOF [about criteria at the July 1993 IETF] was held in a productive atmosphere, but did not achieve what could be called a clear consensus among the assembled attendees. In fact, despite its generally productive spirit, it did more to highlight the lack of a firm direction than to create it.
>
> *(RFC 1994b, p. 2)*

In response to this situation, Gross, chair of the IESG, called for the establishment of an IPng “area”, an ad-hoc constellation of the collection of relevant working groups with a directorate (which he suggested the leaders of himself). At a critical time of escalating diversity, the IESG thus institutionalises a concerting of efforts. The changes in the institutional framework for the design of Internet is elaborated further below.

Returning to the heart of the matters, the contents of solutions and the criteria, there were much variations. The rich and varied set of criteria mirror the fact that many participants in the Internet community felt that they were at a critical point in time, that important and consequential decision had to me made in response to a rapidly changing outside world. Hence, the natural first aim of formulating a tight and orderly set of criteria was not possible:

> This set of criteria originally began as an ordered list, with the goal of ranking the importance of various criteria. Eventually, (...) each criterion was presented without weighting (...)
>
> *(RFC 1994c, p.2)*

> The goal was to provide a yardstick against which the various proposals could be objectively measured to point up their relative strengths and weaknesses. Needless to say, this goal was far too ambitious to actually be achievable (...)
>
> *(SELECT 1992)*

To get a feeling of the kind of considerations, types of arguments and level of reflection about the problem, a small selection of issues are elaborated which related to this paper’s core question of how to make changes to infrastructure technology in order to scale.

Market-orienting Internet

One issue concerned the role and extent market forces, big organisations and user groups should be involved. Of course, none objected to their legitimate role. But exactly how influential these concerns should be was debated. Partly, this issue had to do with the fact that historically the Internet has been dominated by individuals with a primary interest in design. There has until fairly recently not been much attention to the commercial potential of Internet among the community itself. This is clearly changing now (Hinden 1996). The economic and commercial repercussions of Internet was debated as, for instance, the IPDECIDE BOF at the July 1993 IETF confirmed that “IETF decisions now have an enormous potential economic impact on suppliers of equipment and services.” (IPDECIDE 1993). There was widespread agreement that the (near) future would witness a number of influential actors, both in terms of new markets as well as participants in the future development of Internet:

> Remember, we are at the threshold of a market driven environment. (...) Large scale phone companies, international PTTs and such, for example, as they discover that there is enough money in data networking worth their attention. A major point here is that the combination of the IETF and the IAB really has to deliver here, in order to survive.
>
> *(Braun 1992)*

Market forces were recognised to play an important, complementary role:

> [The] potential time frame of transition, coexistence and testing processes will be greatly influenced through the interplay of market forces within the Internet, and that any IPng transition plan should recognize these motivations (...)
>
> *(AREA 1994)*

Still, there was broad consensus that the Internet community should take the lead. At one of the earliest broad, open hearings regarding selection criteria, the IPDECIDE BOF at the July 1993 IETF, it was forcefully stated that “`letting the market decide’ (whatever that may mean) was criticised on several grounds [including the fact that the] decision was too complicated for a rational market-led solution.” (IPDECIDE 1993).

Nevertheless, the increasing tension between the traditional Internet community of designers and commercial interest surfaced. Several pointed out that the Internet designers were not in close enough contact with the “real” world. The “Internet community should not close its eyes to plain reality” (Rekhter and Knopper 1992).

This tension between users, broadly conceived, and designers did not die out. It was repeatedly voiced:

> Concerns were expressed by several service providers that the developers had little appreciation of the real-world networking complexities that transition would force people to cope with.
>
> *(IPDECIDE 1993)*

> More bluntly, I find it rather peculiar to be an end user saying: we end user's desperately need [a certain feature] and then sitting back and hearing non-end-users saying "No you don't".
>
> *(Fleichman 1993)*

Stick or carrot?

Still, the core problem with IPng concerned how large changes could (or ought to) be made, where, how and when to make them — in other words, the transition strategy broadly conceived.

On the one hand, there were good reasons for making substantial changes to IPv4. A number of new services and patterns of use were expected including: real-time, multimedia, Asynchronous Transfer Mode, routing policy and mobile computing. On the other hand, there was the pressure for playing it reasonable safe by focusing on only what was absolutely required, namely solving the addressing space and routing problems. This was recognised as a dilemma:

> There was no consensus about how to resolve this dilemma, since both smooth transition and [new services like for instance] multimedia support are musts.
>
> *(IPDECIDE 1993)*

It was pointed out above that balancing the pressure for changes against the need to protect the installed base is an intrinsic dilemma of infrastructure technology. In the case of IPng, this was amplified by the fact that the core requirements for IPng, namely solving the routing and address space problems, were invisible to most users. They were taken for granted. Hence, there was few incentives for users to change.Why would anyone bother to change to something with little perceived, added value?

In the final version of the selection criteria, addressing this dilemma is used to guide all other requirements:

> [W]e have had two guiding principles. First, IPng must offer an internetwork service akin to that of IPv4, but improved to handle the well-known and widely-understood problems of scaling the Internet architecture to more end-points and an ever increasing range of bandwidths. Second, it must be desirable for users and network managers to upgrade their equipment to support IPng. At a minimum, this second point implies that there must be a straightforward way to transition systems from IPv4 to IPng. But it also strongly suggests that IPng should offer features that IPv4 does not; new features provide a motivation to deploy IPng more quickly.
>
> *(RFC 1994c, pp. 3-4)*

It was argued that the incentives should be easily recognisable for important user groups. Hence, it was pointed out that network operators were so vital that they should be offered tempting features such as controlling “load-shedding and balancing, switching to backup routers” (NGREQS 1994). Similarly, the deep seated aversion for Application Platform Interfaces, that is, tailor-made interfaces for specific platforms, was questioned. Despite the fact that “the IETF does not ‘do’ [Application Platform Interfaces]” (RFC 1995, p. 39), the IESG finally recommends that an exception should be made in the case of IPng. This was because it meets the pressing need for tangible incentives for a transition to IPng (ibid., p.5).

Internet is an infrastructure, not an artifact

A large number of requirements were suggested and debated. They include: topological flexibility, mobile communication, security, architectural simplicity, unique identifiers, risk assessment, network management, variable-length addresses and performance (RFC 1994c). Besides addressing perceived and anticipated requirements, the requirements might have repercussions on the whole infrastructure, not only IPng.

It was repeatedly pointed out that IPng was not only about revising one, self-contained element of the Internet. It was about changing a core element of an infrastructure with tight and oblique coupling to a host of other elements in the infrastructure:

> Matt Mathis pointed out that different proposals may differ in how the pain of deployment is allocated among the levels of the networking food chain (backbones, midlevels, campus nets, end users) (...).
>
> (SELECT 1992)

> I would strongly urge the customer/user community to think about costs, training efforts, and operational impacts of the various proposals and PLEASE contribute those thoughts to the technical process.
>
> *(Crocker 1992)*

This well-developed sense of trying to grasp how one components, here IPng, relates to the surrounding components of the information infrastructure is a principal reason for Internet's success up till now.

New features are included to tempt key users to change. But the drive towards conservatism is linked to one of the most important design principles of Internet, namely to protect the installed base. It is of overriding importance:

> [T]he transition and interoperation aspects of any IPng is *the* key first element, without which any other significant advantage won't be able to be integrated into the user's network environment.
>
> *(e-mail from B. Fink to sipp mailing list, cited by Hinden 1996)*

This appeal for conservatism is repeated ad nauseam. The very first sentence of (RFC 1996) describing the transition mechanisms of IPv6, reads: "The key to a successful IPv6 transition is compatibility with the large installed base of IPv4 hosts and routers" (ibid., p. 1). The pressure for holding back and declining features which might disturb the installed base is tremendous.

"Applying" the principles

A rich and varied set of proposed requirements was worked out. Still, it is not reasonable to hold that the decision was made by simply "applying" the abstract selection criteria to the different proposals for IPng. Despite the fact that the resulting requirements (RFC 1994c) with 17 criteria were "presented without weighting" (ibid., p. 3), a few themes were of overriding importance (IPDECIDE 1993). At this stage, draft requirements had been suggested for more than one year and seven candidates existed but the requirements were "too general to support a defensible choice on the grounds of technical adequacy" and "had so far not gelled enough to eliminate any candidate" (ibid.). The concern for sharper criteria prevailed. It was repeated as late as in March 1994 only two months before the decision was made:

> One important improvement that seemed to have great support from the community was that the requirements should be strengthened and made firmer -- fewer "should allows" and the like and more "musts."
>
> *(AREA 1994)*

The core concern focused on making transition from IPv4 to IPv6 as smooth, simple and uncostly as possible. A few carrots were considered crucial as incentives for a transition, primarily security:

> What is the trade-off between time (getting the protocol done quickly) versus getting autoconfiguration and security into the protocol? Autoconfiguration and security are important carrots to get people to use IPng. The trade-off between making IPng better than IP (so people will use it) versus keeping IPv4 to be as good as it can be.
>
> *(NGDIR 1994)*

Other requirements were to a large extent subordinate or related to these. For instance, autoconfiguration, that is, "plug and play" functionality, may be viewed as an incentive for transition.

The collection of proposed IPng solutions had evolved, joined forces or died. As explained earlier, there was tight interplay between the development of the solutions and the criteria. The real closing down on one solution took place during May-July 1994. In this period, there was extensive e-mail discussions, but more importantly, the IPng Directorate organised a two day retreat 19.-20. May 1994 at BigTen with the aim of evaluating and reworking the proposals (Knopper 1994). Through his and the subsequent IETF in July 1994, an IPng solution was decided upon.

Showdown

By the spring of 1994, three candidates for IPng existed, namely "CATNIP" (evolving from TP/IX), "SIPP" (an alliance between IPAE, SIP and PIP) and "TUBA" (evolving from Simple CLNP). A fourth proposal, Nimrod, was more or less immediately rejected for being too unfinished and too much of a research project.

CATNIP was "to provide common ground between the Internet, OSI, and the Novell protocols" (RFC 1995, p. 12). The basic idea of CATNIP for ensuring this was to have Internet, OSI and Novell transport layer protocols (for instance, TCP, TP4 and SPX) run on to of any of the network layer protocols (IPv4, CLNP, IPX — or CATNIP). The addressing scheme was borrowed from OSI.

A primary objection against CATNIP which surfaced during the BigTen retreat, was that it was not completely specified (Knopper1994; RFC 1995, pp. 14-15). Beyond the obvious problems with evaluating an incomplete proposal, this illustrates a more general point made earlier and illustrated by Alvestrand (1996), area director within IETF: "The way to get something done in the Internet is to work and write down the proposal". Despite appreciation for the "innovative" solution, there was scepticism towards the "complexity of trying to be the union of a number of existing network protocols" (RFC 1995, p. 15).

The TUBA solution was explicitly conservative. Its principal aim was to "minimize the risk associated with the migration to a new IP address space" (ibid., p. 13). This would mean "only replacing IP with CLNP" (ibid., p. 13) and let "existing Internet transport and application protocols continue to operate unchanged, except for the replacement of 32-bit IP[v4] addresses with larger addresses" (ibid., p. 13). CLNP is, as outlined above, OSI's already existing network layer protocol. Hence, the core idea is simply to encapsulated, that is, wrap up, TCP in CLNP packets.

The evaluation of TUBA acknowledged the benefits a solution making use of the "significant deployment of CLNP-routers throughout the Internet" (ibid., p. 16), that is, a solution paying respect to an installed base. Similar to the arguments outlined above regarding the IAB's IPv7 plan to build IPng on CLNP, "[t]here was considerably less agreement that there was significant deployment of CLNP-capable hosts or actual networks running CLNP." (RFC 1995, p. 16). The worries — "including prejudice in a few cases" (ibid., p. 16) — about the prospects of losing control of the Internet by aligning IPng with an OSI protocol were deep-seated.

SIPP was to be "an evolutionary step from IPv4 (...) not (...) a radical step" (ibid., p. 12). SIPP doubles the address size of IP from 32 to 64 bits to support more levels of addressing hierarchy and a much greater number of addressable nodes. SIPP does not, in the same way as CATNIP or TUBA, relate to non-Internet protocols.

The reviews of SIPP were favourable. SIPP was praised for its "aesthetically beautiful protocol well tailored to compactly satisfy today's known network requirement" (ibid., p. 15). It was furthermore pointed out that the SIPP working group had been the most dynamic one in the previous year, producing close to a complete specification.

Still, it was definitely not a satisfactory solution. In particular, the transition plans (based on the encapsulation suggestion originally in IPAE) was viewed as "fatally flawed" (Knopper 1994). A number of reviewers also felt that the routing problems were not really addressed, partly because there was no way deal with topological information and aggregation of information about areas of the network.

In sum, there were significant problems with all three proposals. Because CATNIP was so incomplete, the real choice was between TUBA and SIPP. Following the BigTen evaluation retreat, Deering and Francis (1994), co-chairs of the SIPP working group, summarised the BigTen retreat to the sipp-email list and proposed to build upon suggestions which came out of it. Particularly important, they suggested to “change address size from 8 bytes [=64 bits, the original SIPP proposal] to 16 bytes [=128 bits] (fixed-length)” (ibid.). This increase in address length would buy flexibility to find better solutions for autoconfiguration, more akin to the TUBA solution. These suggestions were accepted by the SIPP working group who submitted the revised SIPP (version 128 bits) to the IPng Directorate together with a new but incomplete transition plan inspired by TUBA. This was accepted in July 1994 as the solution for IPng, finally ready to be put on the ordinary standards track of Internet.

July 1994 - today

Finished at last — or are we?

By the summer of 1994, a recommended candidate for IPng was found. It was called IPv6. It has been put on the standard track (see chapter 4) and made a Proposed Standard in November 1994. One could accordingly be tempted to think that it was all over, that one had found a way which secured the future of Internet. This, however, is not quite the case, not even today. There is a fairly well-founded doubt “whether IPv6 is in fact the right solution to the right problem” (Eidnes 1996). There are two reasons for this, both to be elaborated later:

- There was — and still is — a considerable degree of uncertainty about how to conduct full-scale testing;
- Even if the IPng protocol itself was completed, a number of tightly related issues were still unresolved, most importantly, a transition strategy;

Full-scale testing

A core element of the Internet design principles, what could be said to be the realisation of the Internet pragmatism, is the emphasis on practical experience and testing of any solutions (RFC 1994). Although this principle is universally accepted within the Internet community, the point is that as the installed base of Internet expands, so does the difficulties of actually accomplishing large-scale, realistic testing. So again, how should the principle of realistic testing be implemented for IPng? This worry was voiced fairly early on:

> It is unclear how to prove that any proposal truly scales to a billion nodes. (...) Concern was expressed about the feasibility of conducting reason-

> ably-sized trials of more than one selected protocol and of the confusing signals this would send the market.
>
> *(IPDECIDE 1993)*

The problem of insufficient testing is important because it undermines the possibility of establishing interoperability (ibid.):

> It is also difficult to estimate the time taken to implement, test and then deploy any chosen solution: it was not clear who was best placed to do this.

Current deployment of IPv6 is very slow. Implementations of IPv6 segments, even on an experimental basis, do hardly exist (Eidnes 1996). Even though the phases a standard undergo before becoming a full Internet Standard may be as swift as 10 months, a more realistic projection for IPv6 is 5 years (Alvestrand 1996). The upgrading of IPv6 to a Draft Standard requires testing well beyond what has so far been conducted.

As Internet expands, full-scale testing becomes more cumbersome. Some within the IETF see an increasingly important role for non-commercial actors, for instance, research networks, to function as early test-beds for future Internet Standards (Alvestrand 1996). The US Naval research lab. has implemented an experimental IPv6 segment by June 1. 1996 as part of their internetworking research. The Norwegian research network, which traditionally has been fairly up-front, expects to start deployment of IPv6 during 1997.

Unresolved issues

At the time when the IPng protocol was accepted on the standards track, several crucial issues were still not completed. At the November 1994 IETF immediately following the IPng decision, it was estimated that 10-20 specifications were required (AREA 1994b). Most importantly, a transition strategy was not in place. This illustrates the point made earlier, namely that the actual design decisions are not derived in any straightforward sense from abstract principles. Besides a transition strategy, the security mechanisms related to key management was not — and, indeed, still is not — completed.

A core requirement for IPng was to have a clear transition strategy (RFC 1995). The SIPP (version 128 bits) was accepted as IPng without formally having produced a clear transition strategy because the concerns for facilitating a smooth transition was interwoven with the whole process, as outlined earlier. There was a feeling that it would be feasible to work out the details of the transition mechanisms based on the IPng protocol. It was accordingly decided by the IPng Directorate just prior to the BigTen retreat to separate transition from the protocol.

In response to the lack of a complete transition strategy, informal BOFs (NGTRANS and TACIT) were held at the November 1994 IETF. TACIT was a working group formed during the spring of 1994, NGTRANS was established as a working group shortly after the November 1994 IETF. Both TACIT and NGTRANS were to address the issue of a transition strategy, but with slightly different focus. NGTRANS was to develop and specify the actual, short-term transition mechanisms leaving TACIT to deal with deployment plans and operational policies (NGTRANS 1994). The available time for a transition was to be "complete before IPv4 routing and addressing break down" (Hinden 1996, p. 62). As a result of the deployment of CIDR, it was now estimated that "IPv4 addresses would be depleted around 2008, give or take three years" (AREA 1994b).

From drafts sketched prior to the establishment of NGTRANS and TACIT, the work with the transition strategy was completed to the stage of a RFC only by April 1996 (RFC 1996).

The transition mechanisms evolved gradually. It was early on recognised that a cornerstone of the transition strategy was a "dual-stack" node, that is, host or router. A dual-stack node implements both IPv4 and IPv6 and thus functions as a gateway between IPv4 and IPv6 segments. Dual-stack nodes have the capability to send and receive both IPv4 and IPv6 packets. They enforce no special ordering on the sequence of nodes to be upgraded to IPv6 as dual-stack nodes "can directly interoperate with IPv4 nodes using IPv4 packets, and also directly interoperate with IPv6 nodes using IPv6 packets" (RFC 1996, p. 4).

Progress was also made on closely related elements of an IPv6 infrastructure. The bulk of the IPv4 routing algorithms were reported to be working also for IPv6 routers, a piece of pleasant news in November 1994 (AREA 1994a, p.4).

The additional, key transition mechanism besides dual-stack nodes was IPv6 over IPv4 "tunnelling". This is the encapsulation, or wrapping up, of an IPv6 packet within an IPv4 header in order to carry them across IPv4 segments of the infrastructure. A key element to facilitate this is to assign IPv6 addresses which are compatible to IPv4 addresses in a special way. (The IPv4 compatible IPv6 address has it first 96 bits set to zero and the remaining 32 bits equalling the IPv4 address).

Discussion

Rule following vs. reflective practitioners

A striking aspects of the IPng effort is the difference between abstract design principles and the application of these to situated contexts. A considerable body of literature has, both on a theoretical and an empirical basis, pointed out how human action always involves a significant element of situated interpretations extending well beyond predefined rules, procedures, methods or principles (Suchman 1987). That designers deviate from codified methods and text-books is likewise not news (Curtis, Krasner and Iscoe 1988; Vincenti 1990). Still, the manner deviation from, application to, or expectation from design principles is made the subject of a fairly open and pluralistic discussion is rare. It is not merely the case that the actual design of Internet does not adhere strictly to any design principles. This should not surprise anyone. More surprisingly is the extent to which the situated interpretations of the design principles is openly and explicitly discussed among a significant portion of the community of designers.

When outlining different approaches to systems design or interdisciplinarity, the engineering or technically inclined approach is commonly portrayed as quite narrow-minded (Lyytinen 1987). The Internet community is massively dominated by designers with a background, experience and identity stemming from the technically inclined systems design. The design process of IPng, however, illustrates an impressively high degree of reflection among the designers. It is not at all narrow-minded. As outlined earlier, there are numerous examples of this including crucial ones such as: how the installed base constrains and facilitates further changes, the new role of market forces and the balance between exploring alternatives and closing down.

Aligning actor-networks

The majority of the Internet community has a well-developed sense of what they are designing. They are *not* designing *artefacts* but tightly related collections of artefacts, that is, an *infrastructure*. When changes are called for (and they often are), they do not change isolated elements of the infrastructure. They facilitate a transition of the infrastructure from one state to another.

> Key to understanding the notion of transition and coexistence is the idea that any scheme has associated with it a cost-distribution. That is, some parts of the system are going to be affected more than other parts. Sometimes there will be a lot of changes; sometimes a few. Sometimes the

> changes will be spread out; sometimes they will be concentrated. In order to compare transition schemes, you *must* compare their respective cost-distribution and then balance that against their benefits.
>
> *(Rose 1992b)*

In the vocabulary of actor-network theory (Callon 1991; Latour 1992), this insight corresponds to recognising that the huge actor-network of Internet — the immense installed base of routers, users' experience and practice, backbones, hosts, software and specifications — is well-aligned and to a large extent irreversible. To change it, one must change it into another equally well-aligned actor-network. To do this, only one (or very few) components of the actor-network can be changed at a time. This component then has to be aligned with the rest of the actor-network before anything else can be changed. This gives rise to an alternation over time between stability and change for the various components of the information infrastructure (Hanseth, Monteiro and Hatling 1996).

The crucial but neglected insight of infrastructure design is well developed in the Internet community as the IPng case contains several illustrations of: the difference between short-term and long-term solutions, the debate over CIDR vs. C# and concerns regarding transition mechanisms. The failure to really appreciate this is probably the key reason why the otherwise similar and heavily sponsored OSI efforts have yet to produce anything close to an information infrastructure of Internet's character (Rose 1992). Hanseth, Monteiro and Hatling (1996) compares the OSI and Internet efforts more closely.

An actor-network may become almost impossible to change by having the components accumulating too much irreversibility and becoming too well aligned with each other (Hughes 1983). The components of the actor-network become so to speak locked into one another in a deadly dance where none succeeds to break out. This is not seldom the case with infrastructure technologies. Grindley (1995) describes the collapse of closed operating systems along these lines, without employing the language of actor-network theory. The operating systems were too conservative. They were locked into each other by insisting that new versions were backwards compatible with earlier ones and by tailoring a large family of applications to run only on one operating system. The danger that something similar should happen to Internet is increasing as the infrastructure expands because the "longer it takes to reach a decision, the more costly the process of transition and the more difficult it is to undertake" (IPDECIDE 1993).

Obviously, there are no generic answers to how much one should open an infrastructure technology to further changes, and when to close down on a solution which addresses at least fairly well-understood problems — or simply keeping the old solution without changes for the time being. Internet has pursued and developed what seems a reasonably sound, pragmatic sense of this problem:

> Making a reasonable well-founded decision earlier was preferred over taking longer to decide and allowing major deployment of competing proposals.
>
> *(IPDECIDE 1993)*

Striking a balance between stability and change has to date been fairly successful. Whether this level of openness and willingness to be innovative suffice to meet future challenges remains to be seen. It is anything but obvious.

But what about the future?

The institutionalised framework of Internet is under a tremendous — and a completely new kind of — pressure. This is partly due to the fact that the majority of users come from other sectors than the traditional ones. The crucial challenge is to preserve the relatively pluralistic decision process involving a significant fraction of the community when confronted with situations calling for pragmatic judgement.

> So there it is: politics, compromise, struggle, technical problems to solve, personality clashes to overcome, no guarantee that we'll get the best result, no guarantee that we'll get any result. The worst decision making system in the world except for all the others
>
> *(Smart 1992)*

But only a minority of today's Internet community has acquired the required sense of pragmatism in Internet. There are signs which indicate a growing gulf between the traditional design culture and the more commercially motivated ones (Rekhter and Knopper 1992).

The core institutions of Internet are the IETF, the IESG and the IAB. Despite the fact that the IAB members are appointed from the IETF, the IAB was — especially during the heated debate over the Kobe declaration — poorly aligned with the IESG and the IETF. How, then, can the interests of the IAB seemingly differ so much from those of the IESG and the IETF? I point out a couple of issues I believe are relevant to working out an explanation.

Even if the IAB today is recruited from basically the same population as the IESG and the IETF, this has not always been the case (Kahn 1994). The bulk of the current members of the IAB come from the computer and telecommunication industry (8), two from universities, one from a research institute and one from manufacturing industry. Seven are based in the United States and one from each of Australia, Britain, Canada, the Netherlands and Switzerland (Carpenter 1996). The IAB struggled until fairly recently, however, with a reputation of being too closed (IAB 1990). The minutes of the IAB were not published until 1990. In addition, the IAB was for some time "regarded as a closed body dominated by representatives of the United States Government" rather than the traditional designers of the IETF and the IESG (Carpenter 1996). In connection with the Kobe declaration, this legacy of the IAB was made rhetorically use of and hence kept alive: "Let's face it: in general, these guys [from IAB] do little design, they don't code, they don't deploy, they don't deal with users, etc., etc., etc." (Rose 1992b).

The programmatically stated role of the IAB to advice and stimulate action — rather than direct — has to be constantly adjusted. As Carpenter (1996), the IAB chair, states: "the IAB has often discussed what this means (...) and how to implement it". It seems that the IAB during recent years has become more careful when extending advice in order not to have it misconstrued as a direction. The controversy over the Kobe declaration was an important adjustment of what it is to mean for the IAB to provide advice: "the most important thing about the IAB IPv7 controversy [in the summer of 1992] was *not* to skip CLNP. It was to move the power from the IAB to the IESG and the IETF" (Alvestrand 1996).

The last few years have witnessed a many-folded increase in the IETF attendance, even if it seems to have stabilised during the last year or so. Many important elements of the future Internet, most notably related to Web technology, are developed outside the Internet community in industrial consortia dealing with the HTML protocol family, HTTP, web-browsers and electronic payment. It is not clear that all of the standards these consortia develop will ever get on the Internet standards track. These consortia might decide to keep them proprietary. Still, a key consortium like the WorldWideWeb consortium lead by Tim Berners-Lee has gained widespread respect within the Internet community for the way the standardisation process mimics that of Internet (see http://www.w3.org/pub/WWW). As the organisation of Internet standardisation activities grows, so does the perceived need to introduce more formal, bureaucratic procedures closer to those employed within the OSI: "the IETF might be able to learn from ISO about how to run a large organization: 'mutual cultural infection' might be positive" (IAB 1993).

An important design principle within Internet is the iterative development of standards which combine practical testing and deployment with the standardisation process. This principle is getting increasingly more difficult to meet as the IP revision makes painfully clear. There is a growing danger that the Internet standardisation process may degenerate into a more traditional, specification driven approach.

Non-commercial actors, for instance, research networks, have an important role to play to function as a testbed for future standards (Alvestrand 1996).

Conclusion

To learn about the problems of scaling information infrastructure, we should study Internet. With the escalating use of Internet, making changes required for scaling become increasingly more difficult. Internet has never faced a more challenging task regarding scaling than its revision of IP. After years of hard work most people reckon that IPv6 will enhance further scaling of Internet. But even today, there is a reasonably well-founded doubt about this. We have yet to see documented testing of IPv6 segments.

The real asset of Internet is its institutionalised practise of pragmatically and fairly pluralistically negotiating design issues. Whether this will survive the increasing pressure from new users, interest groups, commercial actors and industrial consortia remains to be seen.

Having argued conceptually for cultivation based strategies to the establishment of information infrastructures (chapter 9) and illustrated an instance of it in some detail in the case of IPv6, we now turn to alternative, more radical approaches to the “design” of information infrastructures. Together with the cultivation based approach, these alternatives approaches make up the repertoire of strategies we have available when establishing information infrastructures.

CHAPTER 11

Changing networks: Strategies and tools

THIS CHAPTER IS IN A PRE-PRE-PRE-DRAFT STATUS

Introduction

Our analysis of information infrastructures have led to a rephrasing of the notion of design. The “design” of an infrastructure is the purposeful intervention and cultivation of an already existing, well-aligned actor-network. This immediately prompts the question of which strategies there exist for such purposeful intervention. One may distinguish between three, generic strategies:

- an evolutionary one: a slow, incremental process where each step is short and conservative;
- a more daring one: a faster process where each step is longer and more daring;
- a radical one: fast changes which are a radical break with the past;

We have conceptually (chapter 9) as well as empirically (chapter 10) elaborated the former of these. The change represents a (limited) number of new features added to the existing ones. The new is just an extension of the existing. Its successful application within Internet is in itself a warrant for its relevance. It is was corresponds to “backwards compatibility” - a well known phenomenon in the world of products (Grindley 1995).

More daring changes imply “jumping” between disconnected networks and involve building a new one from scratch that is unrelated (that is, completely unaligned) to the existing and require *jumping* from one network to another. Examples of this kind of changes is illustrated by e-mail users subscribing to America Online that “jumped” to Internet.

Changing a network through this kind of abrupt changes, however, are often difficult to implement due to the important role of the installed base. Connection to the first network gives access to a large community of communicating partners, while the new one gives initially access to none, making it unattractive to be the first movers. In spite of this fact, this jumping strategy is really the one implicitly assumed in the definition of OSI protocols, and is claimed to be the main explanation of their failure (Stefferud 1994; Hanseth, Monteiro, and Hatling 1996). Such a jumping strategy might, however, be made more realistic if combined with organization and coordination activities. A simple strategy is to decide on a so-called “flag day” where everybody are expected to jump from the old to the new. Still, this strategy requires that the communicating community has a well defined, central authority and that the change is simple enough to be made in one single step. Changing the Norwegian telephone system in 1994 from 6 to 8 digit numbers was done in this way by the Norwegian Telecom who at that time enjoyed a monopoly status.

Radical changes are often advocated, for instance within the business process reengineering (BPR) literature (ref.). Empirically, however, such radical changes of larger networks are rather rare. Hughes (1987) concluded that large networks only change in the chaos of dramatic crises (like the oil crises in the early 70s) or in case of some external shock.

As a strategy for changing information infrastructures of the kind we discuss in this book, relying on abrupt changes is ill-suited and will not be pursued further. Still, the former approach, the evolutionary one, needs to be supplemented. There are important and partly neglected situations where this strategy simply is not sufficient: what we (for lack of a better name) dub “gateway-based” strategies are required for more radical changes. The slow, evolutionary approach involve only modest changes to a well aligned network.

This chapter explains the background for and contents of these, supplementary strategies to the evolutionary approach. We will first illustrate this strategy “in action” through an example, namely the establishment of NORDUnet, a research network in Scandinavia in the 80s. Afterwards, we turn to a more general analysis of the notion of gateways and their role in future information infrastructures.

NORDUnet

Status 83 - 85: networks (and) actors

In the late seventies and early eighties, most Nordic universities started to build computer networks. Different groups at the universities got involved in various international network building efforts. Around 1984 lots of fragmented solutions were in use and the level of use was growing. Obtaining interoperable services between the universities was emerging as desirable - and (technologically) possible.

The networks already in use - including their designers, users, and operating personnel - were influential actors and stakeholders in the design and negotiations of future networks - including Nordunet. We will here briefly describe some.

The EARN network was established based on proprietary IBM technology. The RCSC protocols were used. (Later on these protocols were redesigned and became well known as the SNA protocol suite.) The network was built and operated by IBM. It was connected to BITnet in US. Most large European universities were connected.

The EARN network was linked to the EDP centres. It was based on a “star” topology. The Nordic countries were linked to the European network through a node at the Royal Technical University in Stockholm. In Norway, the central node was located to Trondheim. The network was based on 2.4 - 4.8 KB/second lines in the Nordic countries.

The EARN networks had was used by many groups within the universities in their collaboration with colleagues at other universities. The main services were e-mail and file transfer. A chat service was also used to some extent.

HEP-net was established to support collaboration among physics researchers around the world (?), and in particular among researchers collaborating with CERN outside Basel in Switzerland. This network was based on DECnet protocols. This community represented “big science,” they had lots of money an was a strong and influential group in the discussions about future academic networks. The EDP department at CERN was also very active in developing systems that were established as regular services for this community.

EUnet is (was) a network of Unix computers based on UUCP protocols. EUnet has always been weak in Norway, but was used to some extent in computer science communities. Its main node was located at Kongsberg until 1986 when it was moved to USIT.

EUnet was mostly used by Unix users (doing software development), within academic institutions as well as private IT enterprises.

Norway was the first country outside US linked to ARPANET. A node was set up at NDRE (Norwegian Defence Research Establishment) at Kjeller outside Oslo by Pål Spilling when he returned in 198? from a research visit at.... The second node was established by Tor Sverre Lande at the department of informatics at the University of Oslo in.... This happened when he also returned from a one year (??) research visit at .. Lande brought with him a copy of the Berkeley Unix operating system which included software implementing all ARPANET protocols. The software was installed on a VAX 11/780 computer and linked to ARPANET through a connection to the node at Kjeller. Later on more ARPANET nodes were set up.

NDRE was using the net for research within computer communications in collaboration with ARPA. Lande was working within hardware design, and wanted to use the net to continue the collaboration with the people he visited in US, all using VLSI design software on Unix machines linked to ARPANET.

At that time (??), ARPANET was widely used among computer science researchers in US, and computer science researchers in Norway very much wanted to get access to the same network to strengthen their ties to the US research communities.

At this time Unix was diffusing rapidly. All Unix systems contained the ARPANET protocols, and most Unix computers were in fact communicating using this protocols in the local area networks they were connected to. Accordingly there were lots of isolated IP islands in Norway and the other Nordic countries. By linking these IP islands there would be a huge network.

In Norway the development of one network connecting all universities started in the early eighties. The objective was one network linking every user and providing the same services to all. With this goal at hand, it was felt quite natural to link up with the OSI standardization effort and build a network based on what would come out of that. Those involved tried to set up a X.25 network. First, it was tried build this based on an X.25 product developed by a Spanish company. The quality of this product was low, and it seemed out of reach to get the network up and running.[1] Using this product was given up, and it was replaced by an English product called Camtech. Running DECnet over X.25 was considered. Based on the English product one managed to keep the network running and an e-mail service was established in 84/85 based on the EAN system.

Universal solutions

As the networks described above were growing, the need for communication between users of different networks appeared. And the same was happening "everywhere," leading to a generally acknowledged need for one universal network providing the same universal services to everybody. Such a universal network required universal standards. So far so good - everybody agreed on this. But what the universal standards should look like was another issue.

The was a time of ideologies, and the strongest ideology seems to be ISO/OSI model, protocols and approach. In general there was a religious atmosphere. Everybody agreed that proprietary protocols were bad, and that "open systems" were mandatory. The Americans pushed IP based technologies. They did so because they already had an extensive IP based network running, and extensive experience from the design, operations, and use of this network. The network worked very well (at least compared to others), and lots of application protocols were already developed and in use (ftp, telnet, e-mail,....???).

As the IP based network (Arpanet, later Internet) was growing, the protocols were improved and tuned. New ones were developed as it was discovered that they were urgently needed to make the network work smoothly or new ideas developed as one used the existing services. An example of the first is the development of the Domain Name Service, DNS, mapping symbolic names to digital IP addresses. This service made the network scalable. Further, the decision to build the network

1. The product got the nick-name SPANTAX, after the Spanish airline with the same name. At that time "Spantax" became a almost generic term for low quality services in Norway due to lots of tourists having bad experience with that airline when going to Spanish tourist resorts, combined with one specific event where a flight was close to crash when the pilot thought a large area with lots of football fields was the airport.

on a connectionless transport service made the network flexible, robust, and simple as no management of connections was required during communication sessions.

American research and university communities pushed IP, while both European researchers within the computer communications field and telecom operators pushed ISO. The role of telecom operators had the effect that the whole of OSI is based on telephone thinking.[2] The Europeans wanted a non-IP based solution believing that would close the technological gap between Europe and US.

The OSI idea was.

The IP (Internet) idea was

alliances

links between technology and non-technology, for instance the embedding of telecom operators intentions to expand their monopoly was embedded into the design of X.25 (Abbate 1995).

Nordunet

The Nordunet initiative was taken by the top managers at the EDP centres at the universities in the capitals of the Nordic countries. They had met at least once a year for some time to discuss experiences and ideas. Most of them had a strong belief in computer communication. In Norway the director of the EDP department at the University of Oslo, Rolf Nordhagen, was a strong believer in the importance of computer network technology. He had pushed the development of the BRU network at the university, linking all terminals to all computers. He also worked eagerly for establishing new projects with wider scopes, and he was an important actor in the events leading to the conception of the idea of building a network linking together all Nordic universities. When the idea was accepted, funding was the next issue. The Ministry of the Nordic Council was considered to proper funding organization. They had money, an application was written and funding granted.

Arild Jansen, a former employee at the EDP department in Oslo was now working at the Ministry for Public affairs in Norway and played the role as the bridge between the technical community on the one hand and the political and funding

2. For more on this, see (Abbate, 1995).

communities on the other. He was also the one writing the application for funding. Later he became a member of the steering group.

The idea was one shared network for research and education for all the Nordic countries. This objective almost automatically lead to “openness” as a primary objective. “Openness” was also important for the politicians.

Strategy one: Universal solution, i.e. OSI.

The Nordunet project was established in 1985. Einar Løvdal and Mats Brunell were appointed project coordinators. When the project started, they had hardly the slightest idea about what to do. Just as in the larger computer communications community, those involved in the project easily agreed about the need for a universal solution - agreeing on what this should look like was a different matter.

The people from the EDP centres, having the idea about the project, all believed in the OSI “religion.” Next they made an alliance with public authorities responsible for the field computer networks for research and education would fall into and the funding institution (which was also closely linked to the authorities). Obtaining “universal service” was an important objective for the, accordingly they all supported the ideas behind OSI. These alliance easily agreed that an important element in the strategy was to unify all forces, i.e. enrolling the computer communications researchers into the project. And so happened. As they already were involved in OSI related activities, they were already committed to the “universal solution” objective and the OSI strategy to reach it.

However, products implementing OSI protocols were lacking. So choice of strategy, and in particular short term plans, was not at all obvious. Løvdal was indeed a true believer in the OSI religion. Mats Brunell, on the other hand, believed in EARN. To provide a proper basis for taking decisions a number of studies looking at various alternative technologies for building a Nordic network were carried out:

1. IP and other ARPANET protocols like SMTP (e-mail), ftp, and telnet.
2. Calibux protocols used in the JANET in UK.
3. EAN, an X.400 system developed in Canada.

All these technologies were considered only as possible candidates for intermediate solutions. The main rationale behind the studies was to find best currently available technology. The most important criterion was the number of platforms (computers and operating systems) the protocols could run on.

Neither IP (and the ARPANET) nor the Calibux protocols were found acceptable.The arguments against IP and ARPANET were in general that the technology had all too limited functionality. Ftp had limited functionality compared to OSI's FTAM protocol (and also compared to the Calibux file transfer protocol which FTAM's design to a large extent was based on).The Nordunet project group, in line with the rest of the OSI community, found the IP alternative "ridiculous," considering the technology all to simple and not offering the required services. There were in particular hard discussions about whether the transport level services should be based on connectionoriented or connectionless services[3]. The OSI camp argued that connection oriented services were the most important. IP is based on a connection less datagram service, which the IP camp considered one of the strengths of the ARPANET technology.

JANET was at that time a large and heavily used network linking almost all English universities. The network was based on X.25. In addition it provided e-mail, file transfer, and remote job entry services. The protocols were developed and implemented by academic communities in UK. The fact that this large network was built in UK was to a large extent due to the institution funding UK universities required that the universities bought computers that could run these protocols. JANET was also linked to ARPANET through gateways. The gateways were implemented between service/application protocols. The people involved in the development of the Calibux protocols were also active in and had significant influence on the definition of the OSI protocols. The main argument against Calibux was that the protocols did not run on all required platforms (computers and operating systems).

One important constrain put on the Nordunet project was that the solutions should be developed in close cooperation with similar European activities. This made it almost impossible to go for ARPANET protocols, and also Calibux although they were closer to the OSI protocols unanimously preferred by those building academic networks and doing computer communications research throughout Europe.

The IP camp believed that IP (and the other ARPANET protocols) was the universal solution needed, and that the success of ARPANET had proved this.

The users were not directly involved in the project, but their views were important to make the project legitimate. They were mostly concerned about services. They

3. Connection oriented means.. modeling telephone communication. Connectionless means... modeling ordinary mail (or telegram) services.

want better services - now! But in line with this they also argued that more effort should be put into the extensions and improvements of the networks and services they were using already, and less into the long term objectives. The HEPnet users expressed this most clearly. They were using DECnet protocols and DEC computers (in particular VAX). DEC computers were popular among most Nordic universities, accordingly they argued that a larger DECnet could easily be established and that this would be very useful for large groups. The physicists argued for a DEC solution, so did Norsk Romsenter. Nobody argued for a “clean” DECnet solution as a long term objective.

On the Nordic as well as on the global scene (Abbate 1995) the main fight was between the IP and OSI camps. This fight involved several elements and reached far beyond technical considerations related to computer communications. At all universities there was a fight and deep mistrust between EDP centres and computer science departments. The EDP centres were concerned about delivering (“universal”) services to the whole university as efficient as possible. They thought this best could be done by one shared and centralized services. The computer science departments found most of the time the services provided by the EDP centres as lagging behind and unsatisfactory in relation to their requirements. They saw themselves as rather different from the other departments as computers were their subject. They had different requirements and would be much better off if they were allowed to run their own computers. But the EDP centres were very afraid of losing control if there were any computers outside the domain they ruled.

The computer science departments also disagreed with the EDP centres about what should be in focus when building communication services and networks. The EDP departments focus first on their own territory, then on the neighboring area. This means, first establishing networks across the university. Second, extending and enhancing this so that it becomes linked to the networks at the other universities in Norway, then the Nordic countries. The computer science departments, however, are not interested in communicating with other departments at the same university. They want to communicate and collaborate with fellow researchers at other computer science departments - not primarily in Norway or other Nordic countries, but in US. They wanted Unix computers to run the same software as their colleagues in US, and they wanted connection to ARPANET to communicate with them.

The EDP centres would not support Unix as long as it was not considered feasible as the single, “universal” operating system for the whole university. And they would not support IP for the same reason. And thirdly, they wanted complete control and would not let the computer science department do it by their own either. To get money to buy their own computers, the computer science department had to hide this in applications for funding of research projects within VLSI and other fields. The fight over OSI (X.25) and IP was deeply embedded into networks of people, institutions, and technologies like these ones.

Tor Sverre Lande and Spilling participated in some meetings in the early phase of the project. They were sceptical about Calibux and wanted an IP based solution. They did not have much influence on the Nordunet project and decided to go for the establishment of the IP connections they wanted outside the Nordunet project. And most involved in the Nordunet project were happy when they did not have to deal with the IP camp. At this time there was a war going on between the camps with lots of bad feelings.

As all intermediate solutions were dismissed, it was decided to go directly for an OSI based solution. The first version of the network should be build based on X.25 and the EAN system providing e-mail services. This solution was very expensive, and the project leaders soon realized that it did not scale. X.25 was full of trouble. The problems were mostly related to the fact that the X.25 protocol specification is quite extensive, and accordingly easily leading to incompatible implementations. Computes from several vendors were used within the Nordunet community, and there were several incompatibilities among the vendors' implementations. But maybe more trouble was caused by the fact that lots of parameters have to be set when installing/configuring an X.25 protocol. To make the protocol implementations interoperate smoothly, the parameter setting has to be coordinated. In fact, the protocols required coordination beyond what turned out to be possible.

The project worked on the implementation of the network as specified for about a year or so without any significant progress. The standardization of OSI protocols was also (constantly) discovered to be more difficult and the progress slower than expected, making the long term objectives. The Ministry of the Nordic Council was seriously discussing to stop the project because there was no results. New approaches were desperately needed.

Strategy 2: Intermediate, short-term solutions

At the same time other things happened. IBM wanted to transfer the operations of its EARN network to the universities. Together Einar Løvdal and Mads Brunell then over some time developed the idea to use EARN as backbone of a multi protocol network. They started to realize that OSI would take a long time - one *had to* provide services before that. OSI was all the time ideological important, but one had to be/become more (and more) pragmatic. The idea about "The NORDUNET Plug" was developed. This idea mean that there should be one "plug" common for everybody that would hook up to the Nordunet network. The plug should have 4 "pins:" one for each of the network protocols to be supported: X.25, EARN, DEC, and IP. The idea was presented as if the plug implemented a gateway between all the networks as illustrated by fig.1. That was, however, not the case.

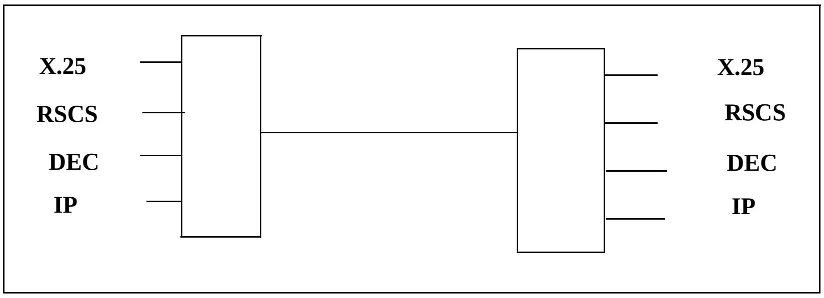

Fig 6. The Nordunet Plug, seen as gateway

The plug only provided access to a shared backbone network as illustrated by fig 2. An IBM computer running EARN/RSCS protocols could communicate only with another computer also running the same protocols. There was no gateway enabling communications between, say, an RSCS and an IP based network.

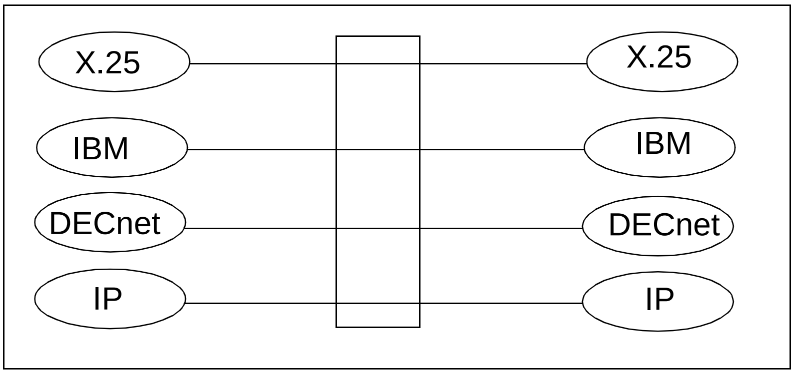

Fig 7. The NORDUNET Plug, as shared backbone

The EARN idea received strong support. Løvdal and Brunell got hold of the EARN lines through a "coup" and the implementation of a Nordic network based on the "Nordunet plug" idea started. They succeeded in finding products making the implementation of the "plug" quite straight forward. First, Vitalink Ethernet bridges were connected to the EARN lines. This means that Nordunet was essentially an Ethernet. To these Vitalink boxes the project linked IP routers, X.25 switches and EARN "routers." For all these protocols there were high quality products available that could be linked to the Vitalink Ethernet bridges.

This solution had implications beyond enabling communication across the backbone. The EARN network was originally designed by a centralized IBM unit and was based on a coherent line structure and network topology. Such coherent topology would be difficult to design by an organization containing so many conflicting interests as the Nordunet project. However, the EARN topology meant that the Nordunet network was designed in a way well prepared for further growth.

Further, the EARN backbone also included a connection to the rest of the global EARN network. A shared Nordic line to ARPANET was established and connected to the central EARN node in Stockholm. 64 KB lines to CERN for HEPnet were also connected.

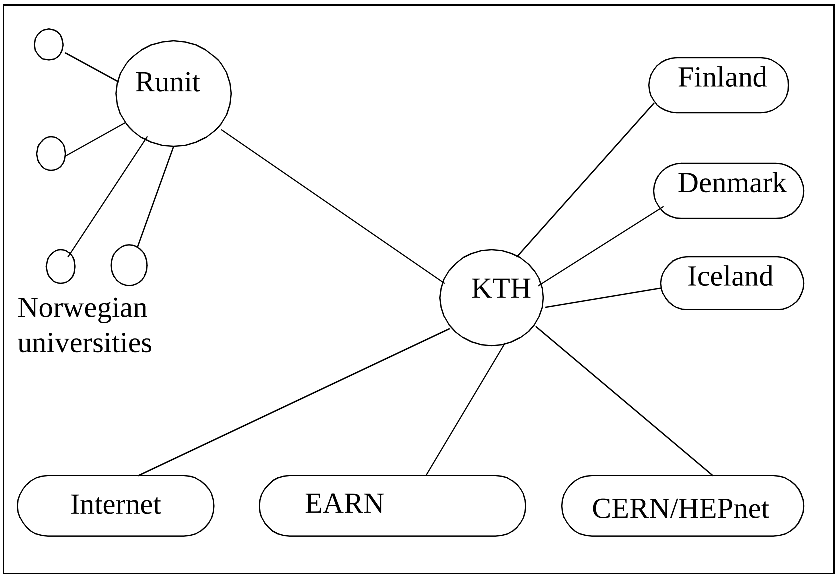

Fig 8. The Nordunet topology

Having established a shared backbone, the important next step was of course the establishment of higher level services like e-mail, file transfer, remote job entry (considered very, very important at that time to share computing resources for number crunching), etc. As most of the networks in use had such services based on proprietary protocols, the task for the Nordunet project was to establish gateways between these. A large activity aiming at exactly that was set up. When gateways at the application level were established, interoperability would be achieved. A gateway at the transport level would do the job if there were products available at the application level (e-mail, file transfer, etc.) on all platforms implementing the same protocols. Such products did not exist.

Before this, users in the Nordic countries used gateways in the US to transfer e-mail between computers running different e-mail systems. That meant that sending an e-mail between two computers standing next to each other, the e-mail had to be transferred across the atlantic, converted by the gateway in the US, and finally transferred back again. The Nordunet established a gateway service between the major e-mail systems used. The service was based on gateways software developed at CERN.

File transfer gateways are difficult to develop as it requires conversion on the fly. CERN had a file transfer protocol, called GIFT (General Interface for FIle Transfer), running on VAX/VMS computers. An operational service was established at CERN. It linked the services developed within the Calibux (Blue book), DECnet, the Internet (ftp), and EARN. The gateway worked very well at CERN. Within Nordunet the Finnish partners were delegated the task of establishing an operational gateway service based on the same software. This effort was, however, given up as the negotiations about conditions for getting access to the software failed.

A close collaboration emerged between the Nordunet project and CERN people. They were “friends in spirit” (“åndsfrender”) - having OSI as the primary long term objective, but at the same time concentrating on delivering operational services to the users.

From intermediate to permanent solution

When the “Nordunet Plug” was in operation, a new situation was created. The network services had to be maintained and operated. Users started to use the network. And users’ experiences and interests had to be accounted for when making decisions about the future changes to the network. Both the maintenance and operation work as well as the use am the network was influenced by the way the network- and in particular the “plug” as its core - was designed. The “plug also became an actor playing a central role in the future of the network.

Most design activities were directed towards minor, but important, necessary improvements of the net that was disclosed its use disclosed. Less resources were left for working on long term issues. However, in the NORDUnet community, this topic was still considered important. And the researchers involved continued their work on OSI protocols and their standardization. The war between IP and X.25 continued. The OSI “priests” believed as strongly as ever that OSI, including X.25, was the ultimate solution. Among these were Bringsrud at the EDP centre i Oslo, Alf Hansen, Olav Kvittem in Trondheim, and Terje Grimstad at NR. Einar Løvdal was fighting for making bridges to IP communities, having meetings with Spilling.

The may be most important task in this period was the definition and implementation of a unified address structure for the whole NORDUnet. This task was carried out successfully.

In parallel with the implementation and early use phase of the “Plug,” Unix diffused fasted in academic institutions, ARPANET was growing fast, ARPANET protocols were implemented on more platforms and created more local IP communities (in LANs), while there was in practical terms no progress within the OSI project.

The increased availability of IP on more platforms led to an increase in use of “dual stack” solutions, i.e. installing more than one protocol stack on a computer linking it to more than one network. Each protocol stack is then used to communicate with specific communities. This phenomenon was in particular common among users of DEC computers. Initially they were using DECnet protocols to communicate with locals or for instance fellow researchers using HEPnet, and IP to communicate with ARPANET users.

The shared backbone, the e-mail gateway, and “dual stack” solutions created a high of interoperability among NORDUnet users. And individual users could, for most purposes, choose which protocols they preferred - they could switch from one to another based on personal preferences. And as IP and ARPANET were diffusing fast, more and more users found it most convenient to use IP. This led to a smooth, unplanned, and uncoordinated transition of the NORDUnet into an IP based network.

One important element behind the rapid growth of the use of IP inside NORDUnet was the fact that ARPANET’s DNS service made it easy to scale up an IP network scalable. In fact, this can be done by just giving a new computer and address and hook it up and enter its address and connection point into DNS. No change is required in the rest of the network. All the network needs to know about the existence of the new node is taken care of by DNS. For this reason, the IP network could grow without requiring any work done by the network operators. And the OSI enthusiasts could not do anything to stop it either.

The coherent network topology and the unified addressing structure implemented also made the network scalable.

Nordunet and Europe

From the very begging, participating in European activities was important for the NORDUnet project. The NORDUnet project also meant that the Nordic countries acted like one actor on the European level. This also helped them coming into an influential position. They were considered a “great power” in line with UK, France and (at that time - West) Germany. However, the relationships change when NORDUnet decided implementing the “plug.” This meant that the project no longer was going for the pure OSI strategy. For this reason the Nordic countries were seen as traitors in Europe.

This made the collaboration difficult for some time. But as OSI continued not to deliver and the pragmatic NORDUnet strategy proved to be very successful, more people got interested in similar pragmatic approaches. The collaboration with the CERN community is already mentioned. Further, the academic network communities in the Netherlands and Switzerland moved towards the same approach.

Throughout its pragmatic strategy and practical success, the NORDUnet had significant influence on what happened in Europe in total. This means that the project contributed in important ways to the diffusion of IP and ARPA/Intenret in Europe - and reduced the possibilities for OSI to succeed.

On the notion of a gateway

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 (Saleh and Jaragh 1998). A gateway in this sense is also called a “converter” and operates by inputting data in one format and converting it 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” (ibid., p. 106).

A gateway, as known from infrastructure technologies for communication and transport, is to translate back and forth between networks which would otherwise be incompatible. A well-known and important example is the AC/DC adapter (Dunn xxx; Hughes 1983). At the turn of the 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 the “battle of systems” unfolded. As a user of electrical lighting, you would have 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 proceeded the AC based one by many years. As described by Hughes (1983) and emphasized by Dunn (198xx), 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.

Potentially confusing perhaps, but we generalise this technically biased notion of a gateway as an artefact that converts between incompatible formats. In line with ANT, we subsequently use “gateway” to denote the coupling or linking of two distinct actor-networks. Compared to the conventional use of the term, our use of the term gateway is a generalization along two dimensions:

- the *coupling* is not restricted to be an artefact but may more generally be an actor-network itself, e.g. a manual work routine;
- the coupling is between *actor-networks*, not only communication networks;

This needs some unpacking to see that it is not just a play with words. To show that this generalized notion of a gateway may actually contribute with anything substantial, we will spell out the roles they play.

Other scholars have developed notions related to this notion of a gateway. Star and Greisemer’s (1992) concept of boundary objects may also be seen as a gateways enabling communication between different communities of practices. The same is the case for Cussins (1996) objectification strategies. These strategies may be seen as constituting different networks, each of them being connected to the networks constituted by the different practices through gateways translating the relevant information according to the needs of the “objectification networks.”

The roles and functions of (generalized) gateways

Generalized gateways (or simply “gateways” from now) fill important roles in a number of situations during all phases of an information infrastructure development. The listing of these roles should be recognised as an analytic vehicle. In practise, a gateway may perform several of these roles simultaneously.

Side-stepping confrontation

The key effect of traditional converters is that they side-step — 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 could co-exist and the decision be delayed after more experience had been acquired.

Side-stepping a confrontation is particularly important during the early phases of an infrastructure development as there are 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.

But side-stepping confrontation is not only vital during the early phases. It is also important in a situation where there already exists a number of alternatives, neither of which are strong enough to “conquer” the others. We illustrate this further below drawing upon e-mail gateways.

When one of the networks are larger than the other, this strategy might be used to buy time to expand and mobilise the smaller network in a fairly sheltered environment. MER???økonomene.....XXXX (David and Bunn 1988)

Modularisation and decomposition

A more neglected role of gateways is the way they support modularisation.The modularisation of an information infrastructure is intimately linked to its heterogeneous character (see chapter 5). As we argued in chapter 8, the impossibility of monolithicly developing an information infrastructure, forces a more patch-like and dynamic approach. In terms of actual design, this entails decomposition and modularization. The role of a gateway, then, is that it encourages this required decomposition by decoupling the efforts of developing the different elements of the infrastructure and only couple them in the end. This allows a maximum of independence and autonomy.

Modularisation, 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 (REFS DIJKSRTEA PARNAS). But the modularisation 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 than one year, the contents is bound to evolve or "drift" (see chapters 5 and 9). This entails that previously unrelated features and functions need to be aligned as a result of this "drifting". The coupling of two (or more) of these might be the result of a highly contingent, techno-economical process, a process which is difficult to design and cater for. Figure XXbyplan illustrates this. Cabel-TV and 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 (XXX) and legislative de-regulation has it become reasonable to link them (REF mansell, mulgan). This gives rise to an *ecology of networks* that later may be linked together by gateways.

Broadbent XXX??? describe an information infrastructure along two dimensions, reach and range, implying that changes can be made along the same dimensions. Changes along the reach dimension amount to adding nodes (or users) to the network while changes to the range amount to adding new functions. Changes to the latter of these, the range, often take place through the drifting and subsequent coupling of two initially, independent networks. Further below we use the case of MIME (Multipurpose Internet Mail Extension, RFC 1341) to highlight this.

Forging compromises

Gateways may play a crucial, political role in forging a compromise in an otherwise locked situation. This is due to the way a gateway may alternative interests to be translated and subsequently inscribed into the same (gateway-based) solution. The key thing is that the initial interests may be faithfully translated and inscribed into one and the same material, namely the gateway. In this way all the alternatives are enrolled and sufficient support is mobilized to stabilise a solution. This is important in dead-lock situations where no alternative is able to "win". Mobilising the support of two or more alternatives through the use of gateways could very well be what it takes to tip the balance.

XX pek ut hvilke av s-in-a sine fem strategier dette minner om

The polyvalent role of gateways

The way a gateway allows interaction with multiple actor-networks makes the context of use more robust in the sense that a user fluently may move between different actor-networks.

???XX HAR VI andre eks enn papir/elektr journal ?????

dual stack, IPng (her en dings)

This polyvalent character of the gateway provides a flexibility that adds to the robustness of the use of the infrastructure.

KNØLETE!!!!!!!!!

Illustrating the roles of gateways

E-mail

E-mail is one of the oldest services provided by Internet. The current version of the standard for e-mail dates back to 1982. That version developed through revisions spanning three years. A separate standard specifying the format of the e-mail message was launched in 1982 together with the protocol itself. An earlier version of formats for e-mail goes back to 1977. The Internet e-mail service consists of two standards, both from 1982: one specifying the format of a single e-mail (RFC 822) and one protocol for the transmitting of e-mails (Simple Mail Transfer Protocol, SMTP, RFC 821).

The e-mail boom in the US proceeded that of Europe and the rest of the world by many years. Already in the 70s, there were a considerable amount of e-mail traffic in the US. There existed several, independent e-mail services in addition to the Internet one, the most important ones being UUCP (a Unix-based e-mail service) and NJE within BITNET (RFC 1506, p. 3). The problem, however, was that all of these were mutually incompatible. Accordingly there was a growing awareness about the need to develop a uniform standard. This recognition spawned CCITT and ISO efforts in working out a shared e-mail standard that could cater for all by providing a "superset of the existing systems" (ibid., p. 3). LITEN KOMMENTAR: prøvde altså opprinnelig på en universlaisme løsning - før man falt tilbake på GW!). These efforts are known as the X.400 standards.

The X.400 initiate enjoyed heavy backing as it aligned and allied with the official, international standardization bodies (recall chapter 4). Especially the lobbying by the West-Germany was influential (ibid., p. 3). Promoting X.400 in Europe made a lot more sense than a corresponding move in the US. This was because the installed base of (Internet and other) e-mail services in Europe was insignificant (see chapter 9). X.400 based e-mail in Europe was fuelled by the free distribution of the EAN e-mail product to research and university institutions.

During the 80s, this created a situation where there where really two candidates for e-mail, namely Internet e-mail and X.400. The large and growing installed base of Internet e-mail in the US (and elsewhere) implied that one would need to live with both for many years to come. After the overwhelming diffusion of Internet the last few years, it is easily forgotten that during the 80s, even the US Department of Defense anticipated a migration to ISO standards. As a result, the Internet community were very eager to develop gateway solutions between the ISO world and the Internet.

>>>> (EMAIL GW rfc1327 kom TO ÅR etter x400(84)!!)

An e-mail gateway between X.400 and Internet has accordingly been perceived as important within Internet. It provides an excellent illustration of the underlying motivation and challenges of gatewaying. Even today, though, “mail gatewaying remains a complicated subject” (RFC 1506, p. 34). The fact that X.400 is really two different standards complicated matters even more. The X.400 from 1984 (written X.400(84)) was originally developed within IFIP Working Group 6.5 and adopted by CCITT. Only in 1988 did CCITT and ISO align their efforts in a revised X.400(88) versions.

The challenge, then, for an e-mail gateway is to receive a mail from one world, translate it into the formats of the other world and send it out again using the routing rules and protocols of that other world. There are two, principal difficulties with this scheme.

First, there is the problem of translating between basically incompatible formats, that is, a thing from one world that simply has no counterpart in the other world. Second, there is the problem of coordinating different, independent e-mail gateways. In principle, e-mail “gatewaying” can only function perfectly if all gateways operate according to the same translation rules, that is, the different gateways need to synchronize and coordinate their operations. We comment on both of these problems in turn.

With X.400(88), an e-mail may be confirmed upon receipt. In other words, it is intended to reassure the sender that the e-mail did indeed reach its destination. This has no corresponding feature in Internet e-mail. Hence, it is impossible to translate. The solution, necessarily imperfect, is to interpret the X.400/ Internet gateway as the final destination, that is, the receipt is generated as the mail reaches the gateway, not the intended recipient. A bigger and more complicated example of essentially incompatibilities between X.400 and Internet is the translation of addresses. This is what in practise is the most important and pressing problem for e-mail gatewaying. This is because the logical structure of the two addressing schemes differ. The details of this we leave out, but interested readers may consult (RFC 1596, pp. 11, 14-29).

The coordination of translation rules for e-mail gatewaying is attempted achieved through the establishment of a special institution, the Message Handling System Co-ordination Service located in Switzerland. This institution registers, updates and distributes translation rules. As far as we know, there exist no survey of the penetration of these rules in currently operating gateways.

MIME

Independent actor-networks that have evolved gradually in different settings, serving different purposes, may need to be aligned and linked through a gateway because they somehow have “drifted” together. An illustration of this is the current developments around e-mail.

The conceptually self-contained function of providing an e-mail service gets increasingly caught up and entangled with an array of previously unrelated issues. An illustration is the current discussion about how to support new applications related to multi-media requiring other kinds of data than just the plain text of an ordinary e-mail message.

Conceptually — as well as historically — e-mail functionality and multi-media file types belong to quite distinct actor-networks. It was anything but obvious or “natural” that the two need to be closer aligned.

There were three underlying reasons why pressure to somehow allow the 1982 version of Internet e-mail to cater for more than unstructured, US-ASCII text e-mails. First and foremost, the growing interest in multi-media application — storing, a manipulating and communication of video, audio, graphics, bit maps, voice — increased the relevance of a decent handling of corresponding file formats for these data types. Secondly, the growth and spreading of Internet prompted the need for a richer alphabet than the US-ASCII. The majority of European languages, for instance, require a richer alphabet. Thirdly, the ISO and CCITT e-mail standard X.400 allows for non-text e-mails. With an increasing concern for smooth X.400/Internet gatewaying, there was a growing need for non-text Internet e-mail.

The problem, of course, was the immense installed base of text-based Internet e-mail (RFC 822). As has always been the Internet policy, "compatibility was always favored over elegance" (RFC 1341, p. 2). The gateway or link between text-based e-mail and multi-media data types and rich alphabets was carefully designed as an extension, not a substitute, for the 1982 e-mail. The designers happily agree that the solution was "ugly" (Alvestrand 1995).

The gateway is MIME, Multipurpose Internet Mail Extension (RFC 1341), and dates back to 1992, ten years after Internet e-mail. What MIME does is fairly straightforward. The relevant information about the multi-media data types included in the e-mail in encoded in US-ASCII. Basically, it adds two fields to the e-mail header: one specifying the data type (from a given set of available options including: video, audio and image) and one specifying the encoding of the data (again, from a given set of encoding rules). Exactly because the need to include different data types in e-mails is recognized to be open-ended, the given list of available options for data types and encoding rules is continuously updated. A specific institution, the Internet Assigned Numbers Authority, keeps a central achieve over these lists.

The political incorrectness of gateways

No one likes gateways. They are regarded as second class citizen that are only tolerated for a little while as they "should be considered as a short to mid-term solution in contrast to the long term solution involving the standardization of network interfaces over which value-added services can be provided" (Saleh and Jaragh 1998, p. 105). A similar sentiment dominates within Internet[4]. An intriguing question — beyond the scope of our analysis — is where these negative reactions are grounded. One reason is that gateways loose information and hence are "imperfect". Infrastructure design, also within Internet, seems to be driven towards "purity" (Eidnes 1996). As this purity is likely to be increasingly difficult to maintain in the future, it would be interesting to investigate more closely into the role of and attitudes towards gateways within Internet.

Based on the experiences outlined in earlier chapters, two kinds of gateways seems to be particularly relevant in health care information infrastructures. One is gateways linking together different heterogeneous transport infrastructures into a seamless web. The other is “dual stack” solutions for using different message formats when communicating with different partners.

E-mail is considered best as carriers of EDI messages. There exist gateways between most available products and protocols. These gateways work fine in most cases. However, they will cause trouble when using features specific for one product or protocol. When using X.400 systems in the way specified by most GOSIPs, saying that the X.400 unique notification mechanisms shall be used, one cannot use gateways between X.400 systems and others not having compatible mechanisms (Hanseth 1996b).

Experiences so far, indicates that implementing and running dual stack solutions is a viable strategy. If a strategy like the one sketched here is followed, implementing tools for “gateway-building” seams to a task of manageable complexity (ibid.).

XXXX en litt mer snappy avslutingig

4. The notion of a gateway is, perhaps surprisingly, not clear. It is used in different ways. In particular, it may be used as a mechanism to implement a transition strategy (Stefferud and Pliskin 1994). It is then crucial that the gateway translates back and forth between two infrastructures in such a way that no information is lost. Dual-stack nodes and “tunneling” (see chapter 10) are illustrations of such gateways. But gateways more generally might loose information as, for instance, the gateway between the ISO X.400 e-mail protocol and the e-mail protocol in Internet. Within the Internet community, however, only gateways of the latter type are referred to as “gateways”. The former type is regarded as a transition mechanism. And it is this latter type of gateways which is not seriously considered within the Internet community.