

# Remembering in the Metaverse: Preservation, Evaluation, and Perception

**Inauguraldissertation**

zur

Erlangung der Würde eines Doktors der Philosophie

vorgelegt der

Philosophisch-Naturwissenschaftlichen Fakultät  
der Universität Basel

von

Florian Müller  
aus Basel (Basel-Stadt)

Basel, 2012



Original document stored on the publication server of the University of Basel:

**[edoc.unibas.ch](http://edoc.unibas.ch)**

The work is licensed under the agreement

"Attribution Non-Commercial No Derivatives - 2.5 Switzerland"

The complete text may be viewed here:

**[creativecommons.org/licenses/by-nc-nd/2.5/ch/deed.en](https://creativecommons.org/licenses/by-nc-nd/2.5/ch/deed.en)**



*Attribution – NonCommercial – NoDerivs 2.5 Switzerland*

*You are free:*



**to Share** — to copy, distribute and transmit the work

*Under the following conditions:*



**Attribution** — You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).



**Noncommercial** — You may not use this work for commercial purposes.



**No Derivative Works** — You may not alter, transform, or build upon this work.

*With the understanding that:*

**Waiver** — Any of the above conditions can be *waived* if you get permission from the copyright holder.

**Public Domain** — Where the work or any of its elements is in the *public domain* under applicable law, that status is in no way affected by the license.

**Other Rights** — In no way are any of the following rights affected by the license:

- Your fair dealing or *fair use* rights, or other applicable copyright exceptions and limitations;
- The author's *moral* rights;
- Rights other persons may have either in the work itself or in how the work is used, such as *publicity* or privacy rights.

**Notice** — For any reuse or distribution, you must make clear to others the license terms of this work. The best way to do this is with a link to the web page <http://creativecommons.org/licenses/by-nc-nd/2.5/ch>.

---

**Disclaimer** — The Commons Deed is not a license. It is simply a handy reference for understanding the Legal Code (the full license) – it is a human-readable expression of some of its key terms. Think of it as the user-friendly interface to the Legal Code beneath. This Deed itself has no legal value, and its contents do not appear in the actual license. Creative Commons is not a law firm and does not provide legal services. Distributing of, displaying of, or linking to this Commons Deed does not create an attorney-client relationship.

---

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät

auf Antrag von

Prof. Dr. Helmar Burkhart  
Prof. Dr. Rudolf Gschwind  
Prof. Dr. Laszlo Böszörményi

Basel, den 13. Dezember 2011

Prof. Dr. Martin Spiess  
Dekan

## Abstract

Electronic memory – computing hardware and software that provides services to extend the capacity of our biological memories – can be seen as the fulfillment of the long-established vision of the MEMEX by Vannevar Bush. In a world of ubiquitous computing, our digital shadows – the proportion of our lives that has some digital representation – is no longer limited to individual documents, but reflects the continuous activities in many parts of our lives. Especially, our digital shadows are no longer isolated, but are connected to other people’s digital shadows in the space of social data and software. Based on three specific case studies, this thesis tries to develop a concept for a future *metaverse archive*: an electronic memory infrastructure that enables the long-term preservation, evaluation and dissemination of the information we acquire throughout our lives.

The first case study focuses on *preservation* and introduces the Permanent Visual Archive (PEVIAR) as a solution to digital preservation. Although electronic storage has become abundant and quite cheap, the long-term preservation of information in the digital realm still poses great challenges. While it is not yet clear whether electronic memory ought to be perfect (in contrast to the benign imperfection of our biological memories), the possibility of safely preserving information in the long term must be given. PEVIAR offers a very specific kind of electronic memory, one that is long-term stable, easily accessible, and authentic, but also very static.

The second case study focuses on the *evaluation* of data. It shows how social data can be used to extract the history of collectives. The email communication of 151 individuals working at the former Enron corporation (amounting to a total of around a quarter of a million of messages) is processed in order to reconstruct, visualize and analyze the social network between these individuals. It will be shown how a physical simulation is suitable for visualizing a very complex network while avoiding information overload and how this simulation not only produces the basis for a suitable visualization, but can further be used to analyze the data in combination with established graph metrics.

The third study focuses on *perception* and shows how context-aware display technologies (more specifically, mixed reality) are an indispensable tool in the capture, evaluation and dissemination of our digital corpora. Since much of the information we acquire is directly related to a real-world context, the recalling and consumption of this information should be able to consider this relation. We focus on spatial context to demonstrate two crucial aspects of context-aware information, namely (spatial) context detection and (spatial) context integration. The concept of hybrid images – images that contain real and virtual parts – is introduced as an example of a context-aware information system applied to the field of architecture visualization.

The three case studies are connected through their role as building blocks for a future electronic memory infrastructure, the metaverse archive. In the conclusion, we summarize the possibilities and limitations of such an archive and highlight some of the societal implications that will need to be addressed.



## Acknowledgements

I would like to thank Prof. Dr. Helmar Burkhart and Prof. Dr. Rudolf Gschwind for having given me the opportunity to be a part of their research groups. During my entire time as a PhD student, they have provided me with guidance, support, inspiration, and confidence. I would also like to thank Prof. Dr. Laszlo Böszörményi for kindly agreeing to act as a co-referee for this thesis.

Throughout my stay at the University, I have had the pleasure of working with and being helped by some of the most interesting, kindest and most intelligent people I have met in my life so far, both from within and outside the university. I would like to thank:

- Dr. Martin Guggisberg, for and with whom I have worked for over a decade in a fruitful manner, and with whom I have shared my office for the last 4 years
- Dr. Peter Fornaro, with whom I had the pleasure to work on the PEVIAR project
- Dr. Tibor Gyalog, with whom I have been able to work on many fascinating projects, and who has once given me very good advice in a very difficult situation
- The entire team of the High Performance and Web Computing group, who are: Prof. Dr. Helmar Burkhart, Prof. Dr. Olaf Schenk, Dr. Martin Guggisberg, Dr. Matthias Christen, Sandra Burri, Robert Frank, Oliver Koch, Phuong Nguyen, Maximilian Riethmann, Sven Rizzotti, Madan Sathe, and Jürg Senn.
- The entire team of the Imaging and Media Lab, who are: Prof. Dr. Rudolf Gschwind, PD Dr. Lukas Rosenthaler, Dr. Peter Fornaro, Dr. Geneviève Dardier, Dr. Simon Margulies, Thomas Angorano, Carl-Christopher Biebow, Daniela Bienz, Sergio Gregorio, Elias Kreyenbühl, Cédric Normand, Patrik Ryf, Tobias Schweizer, Anja-Elena Stepanovic, Ivan Subotic, and Andreas Wassmer
- Jan Torpus for allowing me to participate in a number of fascinating projects
- Peter Mahler and Reto Stibler from the Fachhochschule Nordwestschweiz for introducing me to the world of spatial coordinates and supporting our research efforts in the Lifeclipper2 and HUVis projects
- David Gubler from Fachlabor Gubler AG for his involvement in the PEVIAR project
- Dr. Jürgen Ketterer, Dr. Jean-Noel Gex, and Dr. Christian Neumann from Ilford AG (Marly) for supporting the research efforts in the PEVIAR project
- Willy Tschudin from the University of Basel for supporting our research efforts in the PEVIAR project

Finally, I would like to thank my family and friends for the invaluable support they have provided me with.



# Contents

<b>I</b>	<b>Introduction</b>	<b>9</b>
1.1	Motivation and Outline . . . . .	11
1.2	Electronic Memory . . . . .	12
1.3	From the Internet to the Metaverse . . . . .	17
1.4	The Metaverse Archive . . . . .	22
1.5	Case Studies . . . . .	24
1.5.1	Preservation: PEVIAR . . . . .	24
1.5.2	Evaluation: The Social Graph . . . . .	25
1.5.3	Perception: Mixed Reality . . . . .	25
<b>II</b>	<b>Preservation - The Permanent Visual Archive</b>	<b>27</b>
2.1	Problems of Digital Preservation . . . . .	31
2.1.1	Material Decay . . . . .	33
2.1.2	Hardware Obsolescence . . . . .	39
2.1.3	Software Obsolescence . . . . .	43
2.1.4	Further Problems . . . . .	46
2.1.5	Summary . . . . .	50
2.2	The Permanent Medium Approach . . . . .	51
2.2.1	Ultra-stable carrier . . . . .	51
2.2.2	Visual Interface . . . . .	52
2.2.3	Hybrid Medium . . . . .	54
2.3	Color Microfilm . . . . .	59
2.3.1	Film as an Information Carrier . . . . .	61
2.3.2	Information Capacity of Photographic Materials . . . . .	63
2.4	The PEVIAR Implementation . . . . .	68
2.4.1	Peviar Channel Model . . . . .	70
2.4.2	Modulation Transfer (SFR) . . . . .	71
2.4.3	Granularity (Noise) . . . . .	76
2.4.4	Error-Correction Codes . . . . .	78
2.4.5	Peviar Workflow and Specification . . . . .	80
2.5	Authenticity and Originality in the Digital Archive . . . . .	88
2.5.1	Problems of Authenticity in the Digital Archive . . . . .	89
2.5.2	Cryptographic Techniques . . . . .	90
2.5.3	Peviar: Digital Originals . . . . .	95



<b>III</b>	<b>Evaluation - Harvesting a Social Graph</b>	<b>97</b>
3.1	Social Computing: Theory and Current Practice . . . . .	99
3.2	Social Network Analysis: Computations on Graphs . . . . .	103
3.2.1	Graph Theory . . . . .	103
3.2.2	Questions to Social Networks . . . . .	113
3.3	Introduction to the Enron Email Dataset . . . . .	116
3.3.1	A Brief History of Enron . . . . .	116
3.3.2	Properties of the Dataset . . . . .	120
3.3.3	Related Work on the Enron Dataset . . . . .	125
3.4	A New Approach to Sampling Social Graphs . . . . .	129
3.4.1	Force-directed Layout . . . . .	129
3.4.2	Temporal Aspects of Social Graphs . . . . .	134
3.4.3	Sampling the Distance in Visualization Space . . . . .	137
3.5	System Architecture and Implementation . . . . .	138
3.5.1	System Architecture . . . . .	138
3.5.2	Database and Data Extraction . . . . .	139
3.5.3	Simulation Model . . . . .	142
3.5.4	Weighting and Clustering . . . . .	147
3.5.5	Interactive Visualization . . . . .	153
3.6	Summary . . . . .	159
<b>IV</b>	<b>Perception - Mixed Reality Interfaces</b>	<b>161</b>
4.1	Mixed Reality: Aligning Reality and Virtuality . . . . .	165
4.1.1	Technological Foundations . . . . .	167
4.1.2	Evolution of Hardware Platforms . . . . .	170
4.2	Lifeclipper2: Staging Public Space . . . . .	172
4.2.1	Technical System Implementation . . . . .	173
4.2.2	User Reviews and Lesson Learnt . . . . .	177
4.3	Hybrid Images for Architecture Visualization . . . . .	179
4.3.1	Overall Workflow . . . . .	180
4.3.2	Structure From Motion and Point Cloud Matching . . . . .	182
4.3.3	Model and Virtual View . . . . .	185
4.3.4	Hybrid Image Results . . . . .	187
4.4	Summary . . . . .	189
<b>V</b>	<b>Conclusions</b>	<b>191</b>
5.1	Electronic Memory: Functions and Utility . . . . .	193
5.2	Electronic Memory: First Applications . . . . .	195
5.3	The Metaverse Archive . . . . .	197
5.4	Privacy, Control, and Transparent Citizens . . . . .	201

<b>Bibliography</b>	<b>207</b>
<b>List of Figures</b>	<b>221</b>
<b>List of Tables</b>	<b>222</b>
<b>A Appendix</b>	<b>225</b>
A.1 SFR Measurement . . . . .	225
A.2 RMS Measurement . . . . .	230
A.3 Enron Employee List . . . . .	233
A.4 Curriculum Vitae . . . . .	237

**Part I**

**Introduction**



## 1.1 Motivation and Outline

The developments of the digital revolution – from the early implementations of the von Neumann architecture after the second World War to the smartphones delivering web content in our pockets at broadband speed today – have led to a widespread and dense integration of computing devices and services into our daily lives. The vision of ubiquitous computing – computing that is anywhere, anytime – states that in a future not too distant, computing infrastructure will blend into our real environment, providing an additional layer of informedness as a fabric on top of our world [1]. The emerging infrastructure of an age of information will, in short, make information available when, where and in the manner in which it is required. This thesis is concerned with some of the modalities of this information age. Especially, with the aspect of *electronic memory*. Two parallel developments lead to the significance of this concept. First, the seamless integration of information and computing technology (ICT) in ever more aspects of our lives vastly increases our digital shadow – the portion of our lives that has a digital equivalent. Second, and in direct consequence, the importance of our digital shadow for our everyday lives is growing. Ranging from social networks to personal health monitoring or our collections of letters and photographs, the digital corpora of and about us mean something to us, and they are valuable in the organization and execution of many of our activities. This thesis tries to highlight the meaning and potential of electronic memory through three case studies. It is structured as follows.

In the introduction, several key concepts are explained, namely electronic memory and the conception of the internet as a form of the *metaverse*, a fictional concept introduced almost 20 years ago and more recently gaining popularity among scientists and engineers. Then, the *metaverse archive* is proposed – an infrastructure for electronic memory in the context of the next generation of the Internet. After the introduction, the case studies are introduced as three parts, and can be read independently. In the conclusion, their integration in the concept of the metaverse archive is evaluated. Finally, some of the broader impacts and questions that the described technologies have or will have on our lives are emphasized.

## 1.2 Electronic Memory

In 1945, Vannevar Bush imagined an information system that would allow the effective storage, editing and retrieval of all information encountered throughout a lifetime [2]. He called this imaginary device the *Memex*, short for *memory index*. It consists of a workstation with a storage, transportation and display system for photographic microfilm. While the Memex is often cited as a natural predecessor to the developments in information technology that we witness today, Buckland, in a historical account, puts the work of Bush into perspective [3]. At the time the seminal paper was written, microfilm was already established as an information carrier with very high information density and a long life span (see Section 2.3). While microfilm allowed large quantities of information to be stored in a very limited space, retrieval was still a problem. Early prototypes of microfilm retrieval systems that allowed the search of documents according to some criteria were developed between 1920 and 1930. Bush, while at the Massachusetts Institute of Technology, was involved in a project for the development of a *Microfilm Rapid Selector*, for which a prototype was built between 1938 and 1940. The Memex can be seen as an extension of such a film selector. Apart from the storage and retrieval of film, it also allowed a sort of “active indexing” of the documents. Bush proposed the use of *trails*, associations between documents that the user of the Memex could create as she goes through them, eventually combining several documents in a somehow meaningful sequence. If we take the Memex at face value – that is, as an automated microfilm storage, retrieval and annotation system – it may be inappropriate to attribute the visionary foresight of the information age to Bush. Buckland indicates that the system proposed by Bush was not completely novel. In addition, the introduction of the trails is criticized as a poor alternative to established procedures of indexing developed by documentalists and librarians. However, two ideas that Bush proposed in his article highlight important aspects that the digital revolution has brought about. After some general introductory remarks, he states that a “record, if it is to be useful [...], must be continuously extended, it must be stored, and above all it must be consulted” ([2], p. 39). The emphasis on the use of records, and not only their preservation, results in a great challenge with the ever growing quantity of available records. Bush had witnessed the development of technology that allowed the storage of an entire book on just a few square inches of film. Today, we can store entire libraries on hard disks that have the physical volume of one single paper-back book. Bush, focussing on scientific use of his Memex, imagined that “as the scientist of the future moves about [...], every time he looks at something worthy of the record, he trips the shutter [of his head-mounted camera, the Author] and in it goes” (ibid.). This suggests that in the long run, records will not only be created in great quantities and at increasing frequencies, but the effort required to persist them will decrease considerably. Instead of explicitly creating a document and filling it with content, documents are created automatically at our mere wish.

As we go through life, we amass a large corpus of documents, and at any given time, we should be able to make full use of it.

In contemporary work, the ideas that Bush pioneered are combined in the concept of *electronic memory*. The retrieval system (in the case of Bush, the Microfilm Rapid Selector) is no longer a specific device, but modeled after our *human memory*. In our biological memory, we (more or less effectively) store, extend and retrieve information over an entire lifetime. The ordering of our memories happens in part implicitly, and allows effective and mostly very fast retrieval. Within a split second, we can jump from images of our childhood to what we have had for lunch yesterday, and then again to what we believe to be our most profound philosophical insights regarding the concept of a good life. If we compare this to retrieving documents that we have created on our computers, things look different. Consider a text document that we have written around a year ago. Once we think about it, it will not simply appear on our screen. We will have some knowledge about what the document was about and where we have put it, and if we have an effective system for organizing our documents, we may in fact retrieve it quickly. It may well be, however, that we have a look at our directory trees and cannot quite remember where the document in question is located. We may navigate the directory structure for a while, perform a search for a file name that we think we can remember, or perform a full text search in the hope that the content of the document was indexed by our operating system, and that we enter the appropriate search terms. We may remember that we have sent the document to another person via email, and look for emails to that person from around a year ago. We eventually will find the document, but it will certainly not be as effortless as recalling biological memories. As the amount of digital data of and about us grows, effective retrieval techniques become more important. This is exemplarily demonstrated by an activity usually called *lifelogging*.

Lifelogging is best illustrated by the work of Steve Mann [4]. From on the 1970s, he has been experimenting with wearable computing equipment for every-day video capture under the general term 'personal imaging'. The aim of the project was that individuals using such wearable equipment would at all times have their own *personal information domain* with them. The focus on capturing an individual's visual experience is notable. It is already suggested in Bush's vision, and highlights the aim to move away from a computer-centric perspective of electronic record generation – typing and moving the mouse – towards the integration of automatic record creation in every-day life. Mann has performed his lifelogging activities over decades, and both the miniaturization of wearable computing components and their improvement in performance are evident in his work. At early stages, his equipment consisted of various (heavy) devices worn on his head and attached to his belt, and it allowed him the periodic capture of individual images. In 1994, he was first able to record a live stream

of his visual experience and broadcast it out to the internet in real-time<sup>1</sup>. At the turn of the millennium, his equipment was barely noticeable, being hidden behind a pair of sunglasses. Around that time, Gordon Bell of Microsoft Research had started an ambitious endeavor in the field of lifelogging. Since its beginning in 1999, the *MyLifeBits* project aims at completely recording every aspect of Bell's life in digital form [6] [7]. Apart from historical analogue documents of his life (photographs, letters, faxes, etc., which were digitized) and complete logging data from his computer use (including documents, emails, web activity, and so forth), Bell started to automatically document his life through a neck-worn camera (the Microsoft SenseCam [8]) and audio recordings of conversations and phone calls, coming ever closer what he calls the possibility of *total recall* – the capture of every single aspect and detail in one's life, and the possibility to later recall it precisely.

While enthusiasts like Mann and Bell have put a considerable effort into lifelogging, the ability to 'record our lives' is becoming more and more available. Consider a current smartphone. It is able to capture still and moving images as well as audio, track a users position via various location services, and even determine user activity and context based on measurements of the device's sensors (inertial sensors, compass, audio spectrum, light spectrum). This is to suggest that over time, the main goal of lifelogging - capturing as much information about one's life as possible – will concern every user of ICT technology. It will no longer depend on purchasing appropriate equipment or training oneself to integrate the capture attempt into one's daily routine, but rather be as simple as agreeing to the terms and conditions of a service already pre-installed on our computing devices.

Once the focus moves away from the mastery of capture technologies, the question of how to make use of our extended digital shadow arises. Without any further ado, lifelogs are just a (very) large collection of data. In principle, the problem of electronic record retrieval remains – how do we find the very video that we associate with a certain activity if we do not know the exact time of the event? Certainly, advances in audio and image processing increasingly allow a semantic search of such collections (e.g. searching for classes of shapes or sounds, searching for faces, searching for social constellations, etc.). But a maximization in the amount that is captured may not be the right approach. It is important to consider the value and purpose of the collected data. It is at this point where we ask what to capture, and how to use it that the notion of electronic memory gains relevance. While lifelogging considers the techniques to capture data about our lives, electronic memory is concerned with the access and the utility of this data. And as the name suggests, our electronic memory is modeled after our biological memory – a memory, one should emphasize, that

---

<sup>1</sup>It is reported that his experiment ended in 1996 when visiting the Ecole Polytechnique Federale de Lausanne (EPFL), where at the time the Internet connectivity seems not have been up to his expectations [5]



is very strong in not keeping everything it encounters and in getting rid of quite some of the things it had once stored. Lifelogging systems – or any personal information system – should be designed in a similar manner as our biological memory, which provides us with powerful capabilities. Thus, the functions that we expect our electronic memory to perform are comparable to the functions of our biological memory.

Sellen, a contributor to the MyLifeBits project at Microsoft, has recently given a good account of how the increasing amount of data from and about our lives requires a new perspective on how we manage it [9]. Her main hypothesis is that the knowledge we have about our biological memory is crucial in designing and evaluating electronic memory systems. She proposes that what she calls the *Five Rs*, namely, important functions of our biological memory, should be supported by electronic memory systems if they are to be successful. These functions are *recollection*, *reminiscence*, *retrieval*, *reflection* and *remembering intentions*. Recollection signifies the repeated experience of past memories for the purpose of locating specific information items, such as retracing the activities we have performed at a certain time in order to recall a detail about the situation. When we reminisce, we re-experience past reasons for emotional reasons. This is an activity traditionally supported by artifacts, such as photo albums that help us go back to important moments in our lives, or memorabilia belonging to individuals with whom we share a past. Retrieval is a more general activity, its aim is to locate specific information. It can, but does not have to include recollection. Obviously, it is a very important aspect of our memory, and should be prominent in any electronic memory system. When we reflect on our memory, we try to gain insights into our behavior or the structure of past events. Through reflection, we hope to learn more about ourselves or about our environment, and intend to use that knowledge to make improvements. Remembering intentions, finally, is a function also called *prospective memory*. It allows us to remember to take a certain action in the future based on a past decision. For example, when we make an appointment for the day after tomorrow, we will have to remember to observe it that day.

The five Rs can be considered as *functional* requirements of electronic memory – they state that if we agree to view electronic memory as an extension of our biological memory, it must provide similar functionality. Sellen suggests that these requirements should aid us in determining the usefulness of electronic memory proponents, or, more restrictively applied, provide us with criteria for defining what can be regarded as electronic memory, and what not. It should be noted that the complete fulfillment of all the requirements is not necessary for something to be considered in the domain of electronic memory. If we interpret these functional requirements freely, we could say that electronic memory, generally speaking, can be seen as a biologically inspired metaphor for *personal information systems*. In such a conception, we are close to the concept under which Mann has conducted his life logging experiments. Regarding personal information, we can say that our biological memory is the most important source.

It is also a source that is limited. We cannot hope to remember everything, and to remember everything in complete accuracy. While most would probably agree that this is a feature rather than a bug, the opportunities that a seamless extension of our natural capacities of remembering through the use of computing infrastructure opens up are remarkable.

In conclusion, we understand electronic memory as follows. Electronic memory is a conceptual framework for describing the way in which we structure the information that we acquire with the help of computers. More specifically, it describes services that we expect computers to provide us with based on data collected throughout our daily lives. On a technical level, electronic memory systems consist of at least a mechanism to acquire data, a mechanism to persist, structure and evaluate that data, and on a mechanism to provide us with relevant information based on that data.

### 1.3 From the Internet to the Metaverse

The approach to electronic memory so far has been functional – we have traced some of its technological origins, but mainly formulated expectations about what utility electronic memory is to provide us with. In order to understand the prospects of electronic memory, we must now take a closer look at some of the technologies with which such services could actually be implemented. While ultimately, a range of technologies far too large to be described here will be used, we want to focus on what we believe to be characteristic for the biological metaphor of electronic memory: technologies that are suitable for bridging the gap between the manipulation of artifacts (computers) and immediate experience. Such technologies have been the business of imaginative writers and (pseudo-) prophets for decades. Now that we are witnessing the becoming reality of what was previously only visionary and daunting, it seems suitable to provide a narrative that originates in fiction, but has more recently found its way into reality. We find this in the concept of the metaverse.

The term *metaverse* was popularized in the fictional work *Snow Crash* by Neal Stephenson in 1992 [10]. In his book, he imagines a future world in which the Internet has evolved into the metaverse, a virtual reality into which users log (or rather jack) in on a daily basis. They move about using a visual representation of themselves, called an *avatar*. Users act through their avatar, which is bound by laws similar to the laws of a real environment: it cannot be in two places at once, it can only move at a limited speed, and there are places that are off limits to it. One could say that the user experiences the metaverse mediated by her avatar, but actually, the metaverse is an *immersive* experience. The sensual experience the avatar would have were she a human capable of sensual experience, and the sensual experience the user would have were she an avatar capable of presence in virtual space are one and the same.

While entirely fictional at the beginning (albeit inspired by pioneers of virtual reality), the concept of the metaverse has gained significance in the context of emerging technologies around the Internet in recent years. An illustrative example is that of the Second Life platform [11]. Second Life was brought online in 2003 and was one of the first non-gaming, general-purpose online virtual worlds in which users acted through their avatars. Just as the metaverse, Second Life provides users with a restricted and law-governed virtual environment. Space, which in a virtual world could well be infinite, is limited by coupling it to computing power, a limited real-world resource. The success and impact of Second Life is well documented by the formation and growth of its economy. Second Life users can acquire *Linden dollars*, a currency valid only within the Second Life universe, but convertible into real-world currencies via virtual exchanges [12]. Linden dollars are used as a means of payment for virtual real estate property, services provided in Second Life, virtual objects etc. Around the year 2006, it was reported that the first user of Second Life had become a real-world millionaire through virtual real-estate trades [13]. In 2009, Linden Labs, the owner

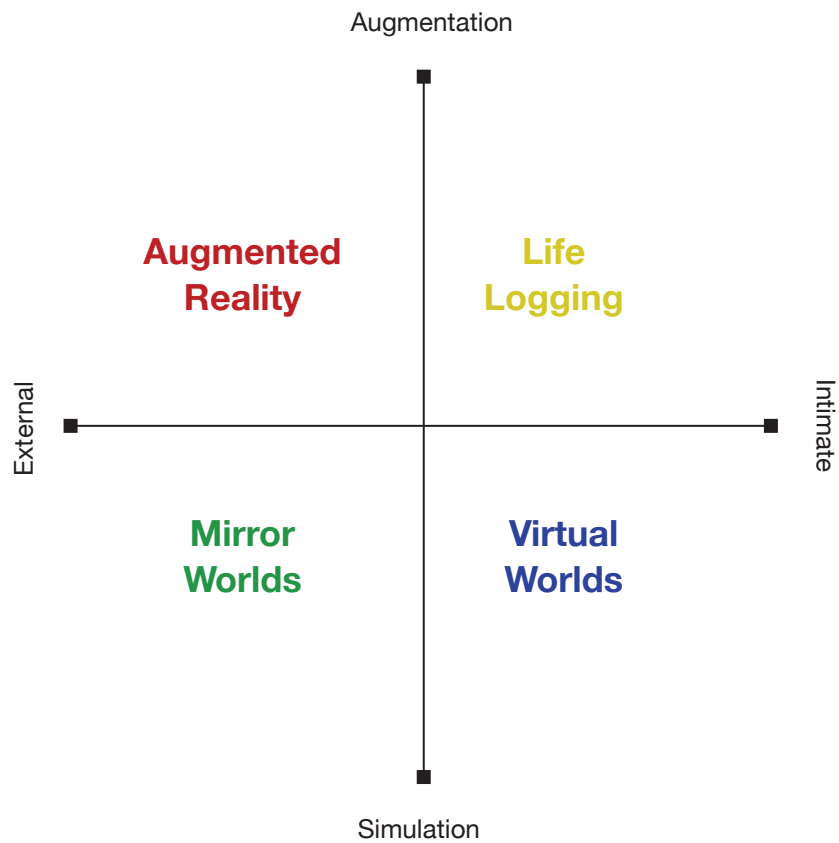
of Second Life, mentioned that 10 individual users in Second Life have annual net earnings of above 1 million US\$, with the most successful user earning 1.7m US\$ in 2009 [14]. The growth and dynamics of Second Life have created a lot of enthusiasm, with universities establishing lecture halls, courses and conferences in Second Life [15], and in 2007, the Maldives narrowly beat Sweden in the race to establish the first embassy of a country in virtual space [16]. While Second Life is not an example of a very immersive virtual community (users do not really feel like they are experiencing their environment with all their senses), it still provided a breakthrough for the wide-spread adoption of virtual reality technologies, including their basic understanding by a wide audience and their introduction as a concept into every-day discourse.

In 2006, the Acceleration Studies Foundation has published a report called *The Metaverse Roadmap* [17]. Based on a broad inquiry among computer industry technologists and academia, in which respondents were asked to assess the future development of various rather specific technological factors (e.g. what percentage of global mobile device users will have broadband internet access from their devices in the year 2016), they synthesized an outline of the future of the Internet. Their principal thesis is that through the continued developments in various technological fields, the Internet will become very much like the globe-spanning virtual world described by Stephenson in *Snow Crash*. The study identifies four crucial technologies and arranges them in a two-dimensional continuum space, which is depicted in Figure 1.

Three of the four technologies (mirror worlds, virtual worlds and augmented reality) can be described as display technologies, while life logging is best described as a technology for capturing and evaluating sensory data in one's life<sup>2</sup>. Mirror worlds are projections of our real world into a space which can, but does not have to be similar to our world. An example would be Google Earth, a three-dimensional interactive model of our planet. Mirror worlds are a projection in that they provide a space for mirroring all sorts of information – geographical, topographical, political, economical, social, etc. – from our real world into a self-contained world which is easy to handle for a user. In a sense, they combine the factuality of our world with the interactivity of state-of-the-art user interfaces, allowing the user to navigate across all scales and in any desired manner. Virtual worlds, on the other hand, are similar to the real world in that they provide a world-like three-dimensional environment in which users navigate in a similar manner as they do in the real world – most importantly, they are usually limited by natural laws such as gravity, friction, and other physical properties that largely determine the possibilities and modalities of interaction. The examples of both Second Life, an existing virtual world, and the metaverse itself, a massively immersive and fictional virtual world, have been introduced in the previous section. If we say that mirror worlds project our real world and that

---

<sup>2</sup>Which, ultimately, will be displayed in some form, possibly using the other three technologies



*Figure 1:* Metaverse continuum space. Four established and emerging technologies are ordered according to two continua External-Intimate and Augmentation-Simulation. The External-Intimate continuum describes the direction of the activities carried out with the help of the technologies: they can either focus on a user's personal and intimate domain, or they can focus on the user's interaction with her physical and social environment. The Augmentation-Simulation continuum distinguishes between providing entirely virtual (simulated, self-contained) services or providing services that integrate virtual elements in a real environment (Image: F. Müller, after [17])

virtual worlds mimic it, we can say that augmented reality (or an *augmented world*) tries to seamlessly integrate the real and the virtual. In augmented reality, the real-time visual perception of a user is modified in such a way that the visual impression of the real environment can be augmented using arbitrary virtual elements which themselves appear to be located in that real environment. Details on the technological foundations of this technology are given in section 4.1.1. Lifelogging, finally, aims at capturing any (or specific) data of a users daily experiences in order to evaluate this data and later use it for some purpose, such as in the context of electronic memory.

The two continua order the four technologies. The axis between simulation and augmentation designates the degree to which the technologies obey to the rules and limitations of our real environment. On the augmentation side, the technologies are fully bound to the user's situatedness in real space and time, augmenting technologies are entirely based on the real environment and try not to control or alter it fundamentally, but rather complement it. On the simulation side, the technologies are free to interpret every aspect of the real environment in any way they desire. They typically adhere to metaphors such as space and time, but have the possibility to adjust them to their own demands. In addition, these metaphors are used in order to provide the user with an intuitive interface, and not because observing them has a value in its own. The axis between external and intimate, on the other hand, describes the direction towards which the technologies are focussed. Intimate technologies are said to focus on the user's individuality, they are relevant for her actions and identity: in the example of virtual worlds, users are provided with an open and possibly limitless space of action and exploration, which they themselves personally conduct. External technologies, on the other hand, are directed towards the world and mostly provide some sort of control over it. In the context of our real environment, by which we are constrained in many ways, they can be seen as an extension of our ability to interact with it: they provide us with more information and increase our range of action.

Taken together, these four technologies show us ways in which the future Internet could develop. In a short-term view, they (especially the ones focusing on display) can be seen as the transition from a two-dimensional web to a *Web3D*. Smart and his co-authors emphasize that the three-dimensional web, i.e. a web that fully embraces 3D graphics technologies, is not a web in which everything is 3D. Rather, it is a *Web 2.5D*, in which some contents such as textual information are still delivered in 2D, but which is capable at any moment to provide a fully three-dimensional and increasingly immersive user experience.

In a middle- to long-term view, and taking into account life logging, the conclusions of the study are more substantial. They claim that the metaverse technologies will increasingly blur the border between our real world and the – or any – virtual world. If we can augment our reality with virtual elements in real time (and in any location), if the virtual places we visit have a correlation in real space, and if activity in the real world echoes in virtuality and vice versa, we can

no longer separate the World Wide Web from the World into which it was built. The result is very close to what Weiser has described as a world of ubiquitous computing. In which ever way we label it, it will be a world in which computers and their services are integrated in nearly every aspect of our lives. By providing interfaces that are very much like the real environment we already know, the new technologies will make this integration appear nearly seamless: the presence of computing infrastructure will be barely noticeable, but the services it provides will be used continuously and extensibly. Since we expect technology to benefit us in the organization of our lives, this computing infrastructure must be highly personalized and aware of the specifics of their users. It is at this point that the concept of electronic memory, and the metaverse archive, gain importance.

## 1.4 The Metaverse Archive

In a negative conception, remembering could mean *not forgetting*, or not losing knowledge about something. Our biological memory is fallible, and over time, we will inevitably forget some things (while others are burnt into our memory). On a social level, we can make our memories redundant by sharing them with others, thereby decreasing the probability that no one will remember a certain event. But even on a social level, things will eventually be forgotten. The use of artifacts – stone, wax tables and papyrus in earlier days; paper, film and electronic storage more recently – as witnesses of past events allows us to maintain knowledge across the boundaries of individual lives and social constellations. The massive surge in electronic storage technology gives us unprecedented possibilities to persist what is and happens in our world.

If we conceive our age as an age of information, storage technologies are of crucial importance. But as has been stated in the introduction of electronic memory, merely preserving data (or information) is not enough. A powerful mnemonic infrastructure must include mechanisms for evaluation and retrieval. In light of the prospects of what has been called the metaverse and the attempts to extend our biological memory with an electronic complement, we can ask ourselves how a future *metaverse archive* would look like. By that, we do not mean one central location where all information is kept, similar to the *Internet archive*, an initiative that aims at progressively and exhaustively keeping track of the development of the Internet as a whole [18]. Rather, we are thinking of a conceptual framework. The metaverse archive is a framework which allows future computing services to make full use of the wealth of information potentially available to them. It is an instance of electronic memory that considers the metaverse technologies as the interface that was previously missing. In this thesis, we present three case studies that we understand to be building blocks of such a metaverse archive. Before we turn to them, we should re-iterate some aspects of electronic memory.

Theoretically, electronic memory is infallible. We can fully control all aspects of how electronic memories are stored and processed (opposed to our brains, which we cannot fully control). The idea is that if we employ the right mechanisms, we will be able to (electronically) remember everything perfectly and forever. This in itself is not trivial, as we will show in the first case study on digital preservation. But given that it would be successful, would we have solved the problem of electronic memory in the metaverse?

Not quite. Not forgetting is only one part of remembering. There are two other crucial components. First, we must be able to distinguish between important and unimportant things to remember. Our memories are not a mere collection of data – we tend to forget unimportant details and only remember what we (consciously or unconsciously) value to some degree. We aggregate and integrate and gain insights from our existing memories, thus creating new memories. Opposed to this, an electronic memory system would theoretically



allow the preservation of all the data we ever acquire. Want and Pering have estimated that the data of an entire life in the form of a full audio and video recording would amount to 100 terabytes [19], a quantity still well above the average storage space we have available as individuals. But the same authors propose that such a capacity could well be common place within a decade. Supposing we in fact keep a perfect record of all our life's data, we must find ways to distinguish meaningful from meaningless data, or, we must find a way to filter a large corpus of data in order to extract useful information. Our real memory constantly does this job for us: of all the impressions we have, we only keep a fraction, namely those seeming relevant. In computer systems, we do not yet have accurate mechanisms for that distinction. It seems safe to just keep all the data, and then later filter it. This function – the aggregation of information from a lot of data – will be one crucial component of a future metaverse archive.

Third, given that we have kept all the information, and are able to distinguish between relevant and irrelevant parts, we lack one more important component: the modalities and contexts of remembering. As has been said, finding specific information in memory is but one function of it. There are other mental activities which are based on our memories, but which do not require the retrieval of specific information, such as reflection and reminiscence. In such activities, we access our memory in a different modality. These modalities must be supported by electronic memory infrastructure. Also, and this may be even harder, we usually access our memory in specific contexts. We can describe this by the following question: how do we know *when to remember what*? Much of our remembering is *triggered* by context: we recall memories when we see something we have seen before, smell something we have smelled before, are somewhere where we have been before etc. These situations can be described as perceptual situations. Electronic memory infrastructure must be able to both detect (or consider) and reproduce perceptual situations. The detection of context serves to identify the contents of the memory that may be suitable given a certain situation, and the reproduction could be valuable as an intuitive interface that provides electronic memories in order to trigger biological memories.

These three components – preservation, evaluation and perception – are all vital components in what we call the metaverse archive. We consider it to be a framework that focuses on using capture, retrieval and interface technologies for purposes of electronic memory. The three case studies presented can be considered as contributions in the field of the individual components. In the conclusion, we will show how they can be integrated into a conceptual view.

## 1.5 Case Studies

Each of the three case studies is concerned with one of the components of the metaverse archive introduced in the previous section. While their impact is integrated in the concept of the metaverse archive, they can also be considered as independent studies that each have their own specific objective.

The first case study on preservation presents the PEVIAR (Permanent Visual Archive) project [20] [21]. It is an attempt to investigate and resolve the principal challenges of digital preservation, i.e. the long-term preservation of information that is represented as digital data.

The second case study on evaluation presents our evaluation of a social graph as given by a data set of corporate email communications [22]. We introduce a graph visualization procedure for social networks that in principle has two properties: first, it allows the visualization and analysis of social networks as they evolve over time, and second, it provides the basis for deriving information about the structural properties of the social network.

The third case study on context presents our work in the field of mixed reality technologies. Mixed reality interfaces are a promising candidate to provide intuitive and effective interfaces for electronic memory recall, mainly because they can be seamlessly integrated into our real environment. Thus, they can operate on the same environment as our biological memory. We exemplarily demonstrate this with two projects: Lifeclipper2 [23] [24], a mobile augmented reality system developed for the use in urban environments and focusing on the (alternate) experience on several time scales, and HUVis (Handheld Urban Visualization) [25], an architecture visualization project that combines several technologies in order to provide users with an interactive tool to visualize their future environments.

### 1.5.1 Preservation: PEVIAR

At a fundamental level, long-term digital preservation faces a problem that can be referred to as the *migration trap*. In essence (and as will be detailed later), any digital archive will have to periodically migrate the information it contains in order to ensure that it is safely preserved. These migrations are not only a source of recurring costs, but also of risk to the collection – in any migration procedure, human error cannot be excluded. The Permanent Visual Archive (PEVIAR) project has developed digital preservation technology that virtually eliminates the necessity for migration.

In addition, it allows the introduction of what we call *digital originals*. In digital storage, the notion of original and copy seems to have been made obsolete, since the contents of any copy of a file are identical. We will show that this circumstance is caused by the immateriality of digital information. Materiality has traditionally been an important source of authenticity and thus originality. In the digital realm, cryptographic techniques try to provide an alternative source

based on complex algorithms. However, apart from their inherent problems regarding reliability, cryptographic approaches introduce a higher risk of loss of information in archives. By binding digital information to material artifacts and making them inseparable, Peviar allows the reconsideration of traditional conceptions of originality and reproducibility in the digital archive.

### **1.5.2 Evaluation: The Social Graph**

Social data – data that reflects the interaction among multiple individuals or organizations – has gained importance since the steep rise of the World Wide Web. It can be used to understand individual and collective behavior, both from an ego-centric and a global perspective. Social data is network data, and the analysis and understanding of such data promises a whole range of interesting services. An early example of this is the success of Google. While they are concerned not with networked individuals, but with networked web pages, they have still shown how the underlying structure of a network can be used to evaluate its members in terms of importance. The *PageRank* algorithm, which they use to prioritize search results, in essence measures the importance of a web page in relation to its location and prominence within a network of web pages. It has revolutionized internet search, or, as it is sometimes enthusiastically said, *solved the search problem*.

While social data is not yet available in the quantities in which web page data is available, it has grown significantly through the adoption of *social media* services. This has a direct consequence for the structure of our digital shadows. They are no longer isolated and confined to a single individual, but include portions of the digital shadow of others. As much as our communication is reflected in our electronic memory, so should the people we communicate with. Our individual memory thus in part consists of a social memory, and the second case study is concerned with the retrieval of information from collective memories as given by social data.

### **1.5.3 Perception: Mixed Reality**

What we consider to be our reality strongly depends on our perception. Mixed reality technologies such as augmented reality provide means to considerably expand the space of possible perceptions. Virtual, computer-generated elements can be made available in a manner that makes them indistinguishable from the real elements of the world around us. This ability to produce perceptual situations which are partly real, partly virtual offers the possibility of novel user interfaces for computing infrastructure. These interfaces are an important building block for the vision of computing that is liberated from bulky interface devices and blends into our world.

The application of mixed reality technologies is demonstrated in two projects. In the Lifeclipper2 project, we took part in the building of a mobile, outdoor

augmented reality system for use in urban environments. It demonstrates how urban settings can be made comprehensible in a historical perspective that looks both into the past and into the future, and how users of such systems can expand their potential to navigate their well-known environment. In the HUVis project, we aimed at providing a specific mixed reality service – the visualization of future architecture – with the restriction that it should not be limited to specific client hardware.

**Part II**

**Preservation - The Permanent Visual  
Archive**



Peviar (Permanent Visual Archive) is a digital preservation system based on a visual information carrier. The motivation for developing it are the fundamental problems observed in digital preservation. Peviar is a radical solution in that it provides a fundamental answer to these problems. It is conceived as a system that is inherently *free from migration*. As will be shown, such a system is not only technically feasible, but also offers interesting properties that other digital preservation systems do not, especially regarding authenticity. Throughout the study, we will use a model for digital preservation, which is introduced here.

In more recent literature, digital preservation has been described as a special form of communication, namely, *communication with the future* [26] [27]. Figure 2 illustrates a simplified and extended version of the detailed model proposed by Moise et. al.<sup>3</sup>. At a certain time ( $t_0$ ), some information is available. It is decided to preserve this information for the future. First, the information is *logically encoded* into a specific representation format, usually called a software format. The logically encoded information is then *serialized* and results in a *bitstream*, a sequence of binary digits (zeros and ones). The bitstream is stored on a physical device, the carrier, for which purpose it is encoded in a carrier-dependent manner. We call this process *physical encoding*. Once it has been stored on the physical device, it is kept for a certain time span in its resulting *physical representation*. At a given time in the future ( $t_0 + \Delta t$ ), the information should be made available. For this purpose, its physical representation is *decoded* (detection and decoding) by a device able to read the carrier. The decoded bitstream is then *deserialized*, i.e. interpreted in a manner suitable for the format that was employed for logical encoding. The deserialized bitstream – the document – is then *rendered* to the intended consumer of the information and made available in a form suitable for use in an information process. Take the example of a word document. You have some information in mind and would like to preserve it. You edit your document using a word processor, thereby logically encoding the information into the word processor format. When you are done, you save the file on your hard disk – i.e. the bitstream of the word document you created is physically encoded for the hard disk and stored on it. From then on, the information is being preserved in the form of its physical representation. At a future point in time, you access your hard drive, and the bitstream of the document is read from the hard drive. The word processor logically decodes it, and presents you with the view of the document. You can now access the information previously stored.

---

<sup>3</sup>The model as proposed by Moise et. al. is based on the OSI-layer architecture for networks. Because we only use it for illustrative purposes, we have simplified it considerably.

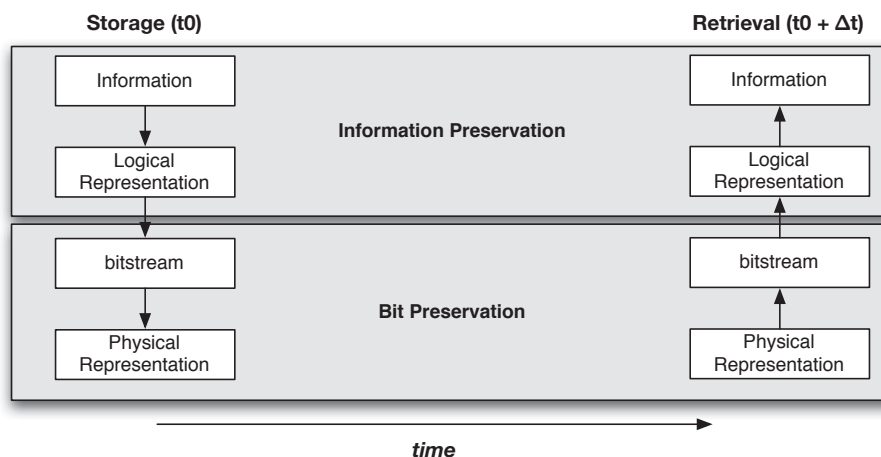


Figure 2: Digital preservation communication model following [27] (Image: F. Müller)

Our model has two layers: bit preservation and information preservation. Bit preservation is responsible for keeping intact and available the bitstream – the sequence of zeros and ones – of the object being preserved (in the OAIS<sup>4</sup> functional model, this would be the task of archival storage). Information preservation, on the other hand, is responsible for keeping the bitstream interpretable, i.e. making sure that we can render the object and gain the information contained in it given the bitstream (preservation planning in the OAIS). Note that bit preservation is possible without information preservation, however, information preservation depends on bit preservation. Note also that there are three time scopes. The present ( $t_0$ ), the future ( $t_0 + \Delta t$ ), and, very importantly, the time between them ( $\Delta t$ ). The model provides us with a systematic view of digital preservation that will help in locating the various problems in their respective place.

The rest of this study is organized as follows. First, the problems of digital preservation are discussed. In section 2.2, the basic strategy of Peviar as a permanent medium is illustrated. Section 2.3 discusses the practical and theoretical foundations of the carrier medium for Peviar, microfilm. In section 2.4, we detail the implementation of the system, including the relevant measurements to determine the technical specification. In section 2.5, we introduce the notion of *digital originals*, which are an interesting candidate to provide authenticity in digital archives.

<sup>4</sup>The Open Archival Information System (OAIS) is a reference model for digital archives, developed by the Consultative Committee for Space Data Systems in 2002 [28] and adopted as an ISO standard in 2003 [29]. Among other things, it provides a functional model for digital archives in which the important stakeholders and technical components and functions are defined.



## 2.1 Problems of Digital Preservation

In 1995, the age of digital computing was already half a century old. The personal computer had been introduced a decade earlier, and the World Wide Web – engineered in 1991 – was on a steady rise. It was only then, however, that the problem of preserving digital information began to attract wider interest. A significant event regarding digital preservation was the publication of Jeff Rothenbergs article *Ensuring the Longevity of Digital Documents* in the Scientific American in 1995 [30]. Even today, it remains one of the most-quoted articles about digital preservation. What Rothenberg explained in his article was that the complex technological chain involved in creating, saving and accessing digital documents makes digital preservation fundamentally different from conventional information preservation. Around that time, the prospects for digital preservation were seen pessimistically. Rothenberg himself noted that “digital information lasts forever – or five years, whichever comes first” ([30], p. 2). Kuny emphasized that “no one understands how to archive digital documents” and that “sustainable solutions to digital preservation problems are not available” ([31], p. 4). He coined the term of the *digital dark ages*, in which he considered the world to be in 1997. Hedstrom warned that digital preservation was “a time bomb for digital libraries” [32]. In the following years, numerous official, academic, and private initiatives have been started to investigate the field of digital preservation, and while new perspectives have been gained on the fundamental problems of digital preservation, no definitive solution has been found.

Let us better understand the problems associated with digital preservation. The aim of digital preservation, as it has been stated, is the safe preservation of digital information in the long term. One could think, then, that it is a problem of electronic storage. Such storage is an important part of our computing infrastructure, and various technologies exist to store vast quantities of information. One application of computer storage is in *backup*. A backup serves to secure documents and computing infrastructure against possible failures, such as device malfunctioning, catastrophic events (fire, flood, etc.) or improper manipulation (e.g. accidental deletion). In a manner, a backup preserves the information contained in documents and information systems in case a loss should occur. One could ask, then, what the difference between a backup and digital preservation is. Or: why is it easy to store and back up digital information, but hard to preserve it? The communication model introduced at the very beginning of this study helps in clarifying this. The relevant difference between digital storage and digital preservation is the time span between storage and retrieval (the size of  $\Delta t$ ). In the case of storage, it is rather short. When  $\Delta t$  goes towards a size that can be considered long-term, we speak of digital preservation. What is considered long-term depends, of course, on the perspective. The OAIS model, for example, defines long-term as a period of time in which technological change significantly impacts a digital archive. We understand long-term to apply to

periods of time that are several decades to centuries long. What can happen in the long term has been intuitively expressed by Rothenberg. He provides a hypothetical scenario to emphasize the challenge of digital preservation ([30], p. 1):

The year is 2045, and my grandchildren (as yet unborn) are exploring the attic of my house (as yet unbought). They find a letter dated 1995 and a CD-ROM (compact disk). The letter claims that the disk contains a document that provides the key to obtaining my fortune (as yet unearned). My grandchildren are understandably excited, but they have never seen a CD before – except in old movies – and even if they can somehow find a suitable disk drive, how will they run the software necessary to interpret the information on the disk? How can they read my obsolete digital document?

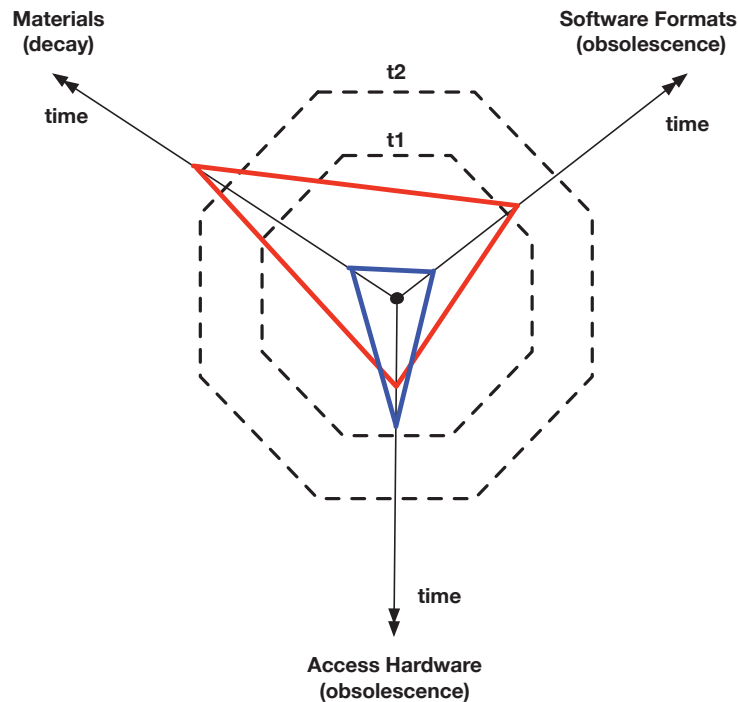
The quotation contains one word crucial for the problem of digital preservation: *obsolete*. The problem of obsolescence<sup>5</sup> concerns both hardware and software. Together with material decay, a factor not mentioned in the scenario, but nonetheless important and discussed later in the article, they are the fundamental hindrance for digital preservation. As will be laid out, there are also other problems associated with digital preservation. Why should these three be considered fundamental? There are two main reasons. First, decay and obsolescence directly affect the technological infrastructure of the digital archive. The impact of other problems on the archive's technical infrastructure, as will be laid out, is not as direct. Second, decay and obsolescence are especially time-dependent. The larger  $\Delta t$ , the more severe their impact becomes. Again, this is not the case for most other problems. Peviar is a digital archiving technology, it aims at providing an infrastructural component for digital archives. Therefore, the three fundamental problems of digital preservation are crucial throughout this thesis. It may well be that other approaches to digital preservation have a different set of focal problems.

Figure 3 illustrates what we call the *problematic triangle*. The core archive infrastructure consists of storage media (materials), access hardware and the employed software formats. As soon as something is stored, the three factors are subject to the impact of time. The state of preservation for any given digital document depends on all three factors, and can be visualized as a triangle. If on any axis, the impact of time is destructive (decay or obsolescence make a document inaccessible), it is lost. The two time horizons depicted by the two octagons illustrate that there may well be several levels of severity for the time-dependant impact. For example,  $t_1$  can be seen as a time horizon in which the factors become problematic, and  $t_2$  as a time horizon in which the factors

---

<sup>5</sup>From latin *obsolescere*, to fall into disuse

become catastrophic. This gradation will be evident in the detailed discussion. Note also that the three axes need not be covariant.



*Figure 3:* The problematic triangle shows that after storage, passing of time inevitably impacts the employed materials (storage media), access hardware and software formats of a digital archive. A gradation of impacts according to time horizons (e.g. problematic / catastrophic) is suggested. The state of preservation of a document is represented as a triangle, in the case of this depiction, the red triangle represents a document that is probably lost, and the blue triangle represents a document that is (currently) well-preserved (Image: F. Müller)

In the remainder of this section, the three fundamental problems as well as additional problems are described.

### 2.1.1 Material Decay

Information storage is always and inevitably based on physical properties of specific materials. These physical properties are not constant – all things material are subject to alteration over time. Within a certain range, alteration of the physical properties used to represent information is tolerable. But when it reaches a certain level, the information may be lost. The manner in which material degradation is related to the retrievability of information is different for analogue and digital information. In the case of analogue information, increasing degradation of the physical properties results in a steady degradation of the

information. Information is lost continuously. In the case of digital information, this is fundamentally different. Information is not lost gradually, but rather not at all, or altogether. This can be referred to as the *digital cliff*: material degradation up to a certain level can be compensated fully, and up to this level, the information is preserved perfectly. If the degradation surpasses this level, the original digital data cannot be reconstructed any more, and the information is lost completely. As an example, we compare the impact of random noise that was added to (a) a photograph of *La gioconda* and (b) a screen capture of a part of the RFC1. The example with the text may seem strange, since it is also based on a perceptible image, and not on non-perceptible digital data. However, for purposes of this illustration, it serves perfectly. Text is digital, and a degradation of the digits either results in them being illegible, or being legible. Either, the information represented by the digits is lost, or it is preserved<sup>6</sup>.

As stated, material decay of carrier media is inevitable, but does not pose a problem up to a certain level. The great majority of storage technologies are based either on magnetic or optical recording. In order to clarify the impact of material decay – or to determine when it becomes problematic – it is discussed exemplarily for some optical and magnetic media. Figure 5 illustrates the level of concern of material decay in the digital preservation communication model. Of the various recording methods (optical, magnetic, magneto-optical, electric etc.), two are discussed exemplarily: optical and magnetic recording. This discussion should aid in understanding the problem of the physical base of recording.

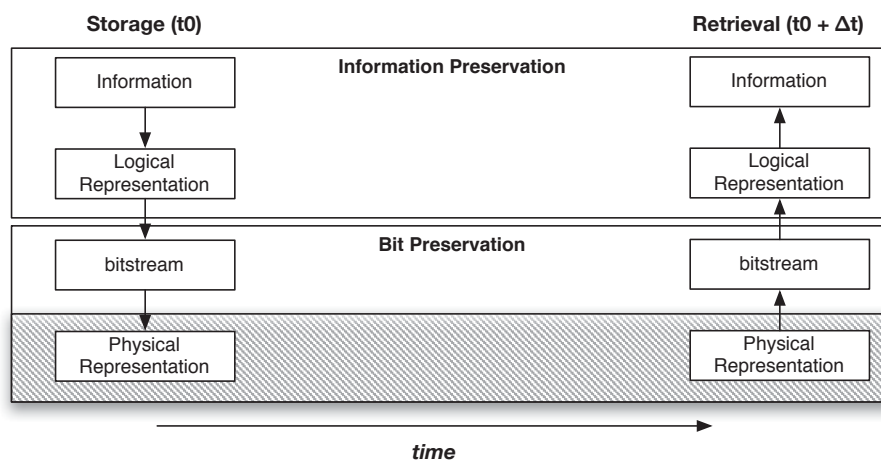


Figure 5: Level of concern of carrier decay in the digital preservation communication model: physical representation (Image: F. Müller)

<sup>6</sup>It is more complicated, since distinct letters are differently affected by degradation regarding legibility. In the case of e.g. a CD-ROM, there would only be two digits (pits and lands)

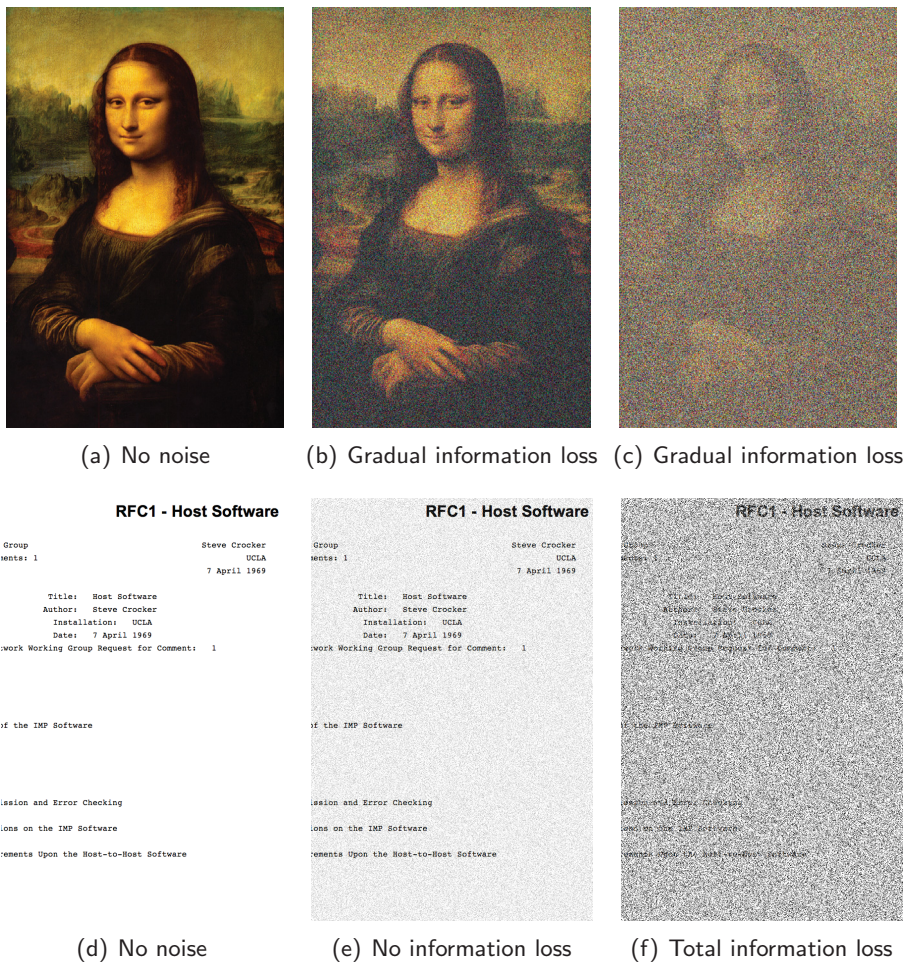


Figure 4: Analogue and digital information degradation (Image: F. Müller)

**Optical recording** From the various optical recording technologies, we cover the most popular disc-based variants. These are read-only compact discs (CD-ROM), once-writable compact discs (CD-R), re-writable compact discs (CD-RW), read-only digital versatile discs (DVD-ROM), once-writable digital versatile discs (DVD-R) and re-writable digital versatile discs (DVD-RW). The operation of all mentioned discs is based on variations in the light-reflecting properties of the disc surface. The light-reflecting properties are varied along one single spiral track. Data is read by rotating the disc and detecting specific light reflection. The manner in which the light-reflecting properties are changed varies considerably from disc type to disc type.

- *Read-only discs (CD-/DVD-ROM)* – these discs are pressed. The change in light reflection results from changes in distance between the laser beam

used for detection and the point of reflection. This is achieved by structuring the track into *pits* and *lands*, where pits are closer to the detector than lands.

- *Write-once discs (CD-/DVD-R)* – these discs are burnt by a laser in a CD-/DVD-R-drive. Discs have a layer of dye above the reflective surface. The light reflectance change is achieved by selectively burning and thereby removing parts of the dye with a laser, creating a spatial structure similar to the pits and lands.
- *Re-writable discs (CD-/DVD-RW)* – these discs are burnt by a laser in a CD-/DVD-RW-drive. Discs have a phase-change layer, i.e. a layer comprised of a material which is able to change its phase locally (between stable-crystalline and metastable-amorphous). The light reflectance change is achieved by selectively heating up the phase-change layer, thereby locally changing the phase on the track and creating a reflectance structure similar to the pits and lands.

The lifetime of optical media depends on how stable the physical properties governing light reflection are. In the case of read-only media, this is the stability of the pressed reflective layer, which is usually aluminum. In the case of write-once and rewritable media, this is the stability of the dye or phase-change layer and of the reflective layer. In addition, the lifetime depends on resistance against scratches. In recent research on the longevity of once-writable optical media (which, among the three types, is the one most often used in archives), it seems that the dye properties are the most important ones [33]. It was shown that phtalocyanine-based dyes combined with a gold-silver alloy reflective layer provide the best stability. The case of once-writable DVD is not as simple: too little is known about the composition of the dyes. Their composition is not necessarily published by the manufacturers, and since the DVD technologies are still in intense development, the composition is subject to change in the process of optimization. While some manufacturers claim very high lifetime for their products (Kodak 300 years gold archival disc), these numbers should be treated with great care. Slattery concludes that CD-R and DVD-R *can* in fact be very stable, meaning that data is guaranteed to be safe for several decades. It is also noted, however, that distinguishing archive-grade media from non-archive-grade on the complex CD/DVD market is difficult, especially for users with no professional expertise in the field. In general, CD and DVD can be considered as mass products in a market with strong competition. Low-priced media often have poor quality. Signs of decay such as delamination and stains can often be observed after only several years. The effects of not only material decay, but also the quality of the write process of CD and DVD are reported by Bienz [34] for the case of a large corporate research archive. Additional factors such as labeling have been incorporated in a study by Youket and Olson [35]. They report on a natural and two artificial ageing studies of compact disc service life.

Their work elaborates the multiple factors involved in media degradation and provides the basis for appropriate archival strategies.

**Magnetic recording** The most common magnetic storage technologies are hard disk drives and magnetic tape. Magnetic tape is – at the moment – the cheapest technology in terms of amount of storage per price (see Table 1). Traditionally, magnetic tape has been used for backup and archiving of digital data: not only because tape is cheaper, but also because it is considered more long-term reliable. This relation has recently come into question. Hard-disk based archives have started to become attractive alternatives to tape archives, mainly because their price has dropped considerably and because disk-based archives are much faster or, in other words, not as off-line as tape.

Technology	Class	Price/GB	Relative Price
LTO-4 Magnetic Tape	tertiary	0.02	1.0
Harddisk	secondary	0.11	5.5
DVD+R dual layer	secondary	0.15	7.5
CD-R	secondary	0.56	28
Solid State Disk	secondary	1.49	74.5
RAM (DDR)	primary	9.98	499.0

*Table 1:* Comparison of various storage technologies regarding their cost. *Price/GB* is the absolute price in Swiss Francs, *Relative Price* is the price compared to LTO-4 magnetic tape. Calculations are based on prices found at an online electronic equipment reseller in October 2011 (Source: [36])

Hard disks and tapes use magnetic recording technologies. Instead of applying visual markings to a material as in the case of optical media, the magnetization of a material is changed locally. The application and detection of the magnetization is achieved using a so-called *read-write head* (head). Hard disks store data on one or several platters. These platters are mounted on a spindle, which holds a motor that allows the rotation of the platters at very high speeds. For reading and writing data, the head is positioned over the rotating platters and either detects the magnetization at a specified location (read) or changes it (write). Tapes store data on a magnetic emulsion (e.g. iron particles in polyurethane) applied to a thin polyester film base. Data is read and written by moving the tape head over the tape and either detecting or inducing magnetization. Several magnetized tracks are applied to the tape, either in linear or in helical layout. Compared to optical discs, both hard disks and magnetic tape are not only affected by decomposition of the materials they consist of, but also by mechanical fatigue.

**Summary** A stable medium is not sufficient for successful preservation of information, but it is the first prerequisite. If the carrier material fails, all other

aspects are irrelevant. There are two important lessons from the treatment of digital storage materials. First, compared to the number one carrier of the “analogue world”, paper, the carriers of the “digital world” are far inferior. Also, the involved technologies are far more complex. Second, there are data carriers that are stable over a time period suitable for use in archives. But it is important to know that not all media fulfill this criterion. The choice of a suitable carrier must be careful and well-informed. As evident from the discussion, it is very hard to provide adequate lifetime estimates for storage media. One reason for this is that digital storage has not been around for much time compared to books – it has been introduced a little over 50 years ago. Also, the replacement of old technologies through more recent ones limits the time a specific technology is in service. These two factors combined result in the lack of a sufficient empirical base for estimating average material lifetime based on natural ageing. Take the example of punch cards. Today, many punch cards that have been created several decades ago still exist. One could say that they provide a sufficient empirical base for studying the natural ageing for periods of several decades. The problem is that today, no one is interested in the ageing properties of punch cards – they have long been replaced by other technologies.

It would nonetheless be advantageous to be able to capture the lifetime of different storage media in years: archivists and preservationists could rely on these numbers to plan their preservation strategy. Such numbers are indicated by various sources, but since they are highly questionable in their veracity, they offer little consolation. Rothenberg estimates the lifetime of optical media at 5-59 years, of digital tape at 2-30 years, and of magnetic discs at 5-10 years. Kodak, a manufacturer of so-called archival-grade CD-R, states that “the lifetime of Kodak Photo CD, and Kodak Writable CD Media with InfoGuard Protection System, under normal storage conditions in an office or home environment, should be 100 years or more” [37]. Byers [38] mentions manufacturer-conducted experiments that suggest that once-writable optical media last for 100 to 200 years, and rewritable optical media last for 20 to 100 years. However, he also suggests that both the experimental setup (accelerated ageing, see Section 2.3) and the conclusions from the measured data leave much room for interpretation [38]. In a natural ageing study at the Library of Congress, Manns and Shahani have observed only slight degradation over a 7 year time span for a sample size of 125 CD-ROMs [39] containing digital audio recordings<sup>7</sup>. Nestor, a german network for digital preservation, reiterates that manufacturer lifetime estimates may not be reliable. They do not provide an estimate of their own, however, they characterize storage media with respect to their suitability for long-term archiving. While the magnetic diskette is considered *unsuitable*, magnetic discs, magnetic tape and optical media are all considered *conditionally suitable* (no technology is considered suitable, or even well suitable [40]). In face of the

---

<sup>7</sup>The empirical base for observing natural ageing spans, as of 2011, a little over a decade, which is a very narrow time horizon for long-term preservation



uncertainty of the permanence of storage media, Bradley concludes that “the question of how to build a permanent carrier was never really answered” ([41], p. 153). In an article about the reliability of hard disks, Rosenthal concludes that safely storing bits for long times is an unsolved problem [42]. In regard of the limitations and uncertainties associated with carrier media, he suggests that the safe preservation of documents must be conceived with the greatest possible independence from any single medium. Instead, it can be seen as an attempt to maximize the following three parameters [42]:

1. The number of copies of a document
2. The number of independent copies of a document
3. The frequency with which the copies of a document are audited (proof-reading)

The first point accounts for the possible failure of any one single medium. The second point emphasizes the importance of independence between copies. Independence can be technological, such as keeping one copy on a hard disk and one copy on a tape. Independence can also be spatial: instead of keeping all copies at the same location, they are spatially distributed. The final point is crucial. Since no reliable lifetime estimates for media are available, the media must be audited as often as possible. Material degradation is steady, and the more often media are audited, the more likely it is that increasing and threatening material degradation will be detected early enough in order to take countermeasures. While Rosenthal’s advice of redundancy and audition may seem trivial, it is the best way to cope with the limitations and uncertainties of storage media. Whether media have to be replaced every five, ten or twenty years – there is no way to really know in advance. The only viable strategy are a careful evaluation of available technologies, redundant storage, and regular auditions and, if necessary, replacements.

### **2.1.2 Hardware Obsolescence**

The computer industry is known for its fast innovation cycles. The invention of new technologies and the optimization of production capabilities have enabled steady and exponential progress over the past decades. This is most notably expressed in what is called Moore’s law. In its original version of 1965 [43], it predicted that the number of transistors in integrated circuits would double every year. Some decades later, it has been shown that this number in fact doubles about every second year. Rapidly increasing capacity is not a phenomenon exclusive to CPUs, for which it is most prominently known, but can be observed regarding other performance characteristics of computing equipment such as network bandwidth or storage capacity [17]. In the case of storage, it is not only the capacity that is changing – this would not be a problem. But it is the technologies that are used to provide it. Figure 6 illustrates this: it is a

photograph of various storage technologies that have been in use at some point in the last 20 years, among them various magnetic diskette, magnetic tape, and optical disc technologies.



Figure 6: Various storage technologies (Image: R. Gschwind)

Consider the case of affordable external storage for the end-user. For a long time, floppy diskettes have been used, providing only a capacity of 1.44 megabytes in the most popular variant. In 1994, the Iomega ZIP drive was introduced, which provided an initial capacity of 100 megabyte. Iomega soon produced a 250 megabyte variant of the disk. However, around the year 2000, CD recorders came into a low price range, and subsequently became the dominant technology for external storage at home, providing 700 megabyte of storage apiece. Today, they are paralleled by DVD recorders, which provide a greater capacity (4 gigabyte apiece). The rapid succession of storage technologies and their incompatibility among one another and, over time, to the computing systems of which they are a peripheral part is called hardware obsolescence. Data access hardware is the bridge between the physical representation of the bitstream on the carrier and the software that is responsible for interpreting and rendering the bitstream. If it fails, the bits may well be preserved physically, but are inaccessible. In the digital preservation communication model, hardware obsolescence concerns bitstream preservation, or more precisely, bitstream accessibility (see Figure 8).

Hardware obsolescence even applies within a technological family. Take the example of magnetic tapes. Manufacturers are constantly increasing tape capacities in order to be able to support the ever growing storage space requirement. New tape technologies are not always compatible among one another.

Figure 7 gives an overview of the compatibility between various products of the tape technology manufacturer Quantum. Each row shows the compatibility of a certain tape medium recorded with a specific format<sup>8</sup> to the various tape drive technologies provided by Quantum. Compatibility is either given in read and write mode (rw), in read mode only (r), or not at all (-). Some drives are backwards compatible, i.e. they support media that were used with older drive technologies, especially in the DLT 2000 to DLT 8000 series. In other series, the compatibility is read-only. Media are never backwards compatible (all entries below the diagonal are incompatibilities). When long-term preservation is considered, this chart of tape technologies coming from the same manufacturer over a period of 13 years shows that obsolescence is a matter of rather short time.

Tape Medium	Recorded Format	Size GB	Tape drives												
			DLT 2000	DLT 2000 XT	DLT 4000	DLT 7000	DLT 8000	DLT1 VS80	DLT VS 160	DLT V4	SDLT 220	SDLT 320	SDLT 600	SDLT 600A	DLT S4
			1993	1995	1994	1996	1999	2001	2003	2005	1998	2002	2004	2005	2006
DLTape III	2000	10	rw	rw	rw	rw	rw	-	-	-	-	-	-	-	
DLTape IIIXT	2000XT	15	-	rw	rw	rw	rw	-	-	-	-	-	-	-	
DLTape IV	4000	20	-	-	rw	rw	rw	r	-	-	r	r	-	-	
	7000	35	-	-	-	rw	rw	-	-	-	r	r	-	-	
	8000	40	-	-	-	-	rw	-	-	-	r	r	-	-	
	VS80 & 1	40	-	-	-	-	-	rw	r	r	r	r	-	-	
DLTape VS1	VS 160	80	-	-	-	-	-	-	rw	r	-	-	r	-	
	V4	160	-	-	-	-	-	-	-	rw	-	-	-	-	
SDLT I	220	110	-	-	-	-	-	-	-	-	rw	rw	r	-	
	320	160	-	-	-	-	-	-	-	-	rw	r	-	r	
SDLT II	600	300	-	-	-	-	-	-	-	-	-	-	rw	-	
	600A	300	-	-	-	-	-	-	-	-	-	-	-	rw	
DLTape S4	S4	800	-	-	-	-	-	-	-	-	-	-	-	rw	

Figure 7: Compatibility chart for Quantum DLT tape drives, media, and recording modes. Compatibility is either not given (-), read-only (r) or read-write (rw) (Data: [44])

Regarding the effective impact of obsolescence, one final example is provided. It was reported by Gregorio [45]. In the course of a color reconstruction of old (1955), faded color photographs in 1991, the Imaging and Media Lab (University of Basel) produced 35mm slides from 6 by 6 inch color slides for the Basel Cultures Museum. Fourteen years later, in 2005, the museum asked whether the original scans of the slides still existed and whether they could be delivered to them. As it turned out, the scans existed on a DAT-DDS1 tape. However, the current drives used (DDS4) were not compatible with DAT-DDS1. A DAT-DDS1 drive was found on Ebay. After overcoming some problems with drivers and interconnect cables (SCSI1), the drive was operational, and the data could be read. It was discovered that the scans were stored in a proprietary format (“TIFF did not exist in 1991!” , [45], p. 92). The old Fortran programs used to process the scans were analyzed, and a C program was written to convert the old proprietary format to TIFF. Finally, the original scans could be delivered to the museum – eight months after their inquiry, and with a lot of luck that all

<sup>8</sup>The chart was compiled by Fujifilm, a manufacturer of tape media, hence the focus on media rather than drives

the components still existed.

This example shows how real the problem of hardware obsolescence can become. Also, it shows that the consequences of obsolescence are not necessarily catastrophic. If a technology becomes obsolete, it may still be accessible for quite some time. Obsolete hardware may still be available for purchase or in the form of spare parts, and some way for connecting them to current computing hardware will be found. However, it may be that the solution to the problem is time-consuming and very expensive (see Section 2.1.4 for economic considerations). Also, the longer the time period in which no action is taken, the higher the probability that the data cannot be read any more.

When it comes to estimating typical obsolescence cycles, no consensus has been established. Some authors are very pessimistic regarding the time span before a hardware technology becomes obsolete: Rothenberg estimates it at five years [30], Kuny at (improbable) 18 months [31]. The examples mentioned in this section have shown that technologies can fall into disuse within only several years. On the other hand, there are examples of storage and related technologies that have proven to be quite persistent. The Universal Serial Bus (USB) was introduced in 1994, and the current second version of the standard is backwards-compatible, meaning that USB-attached storage devices such as external hard disks bought in 1994 have not become obsolete for at least 16 years. The 3.5 inch magnetic diskette (high density, capacity 1.44 megabyte) has been introduced in 1986, and both drives and media can still be bought today, meaning that the technology has been in use for over 20 years. The low numbers provided by some authors are based, of course, on contrary examples, i.e. technologies that have become obsolete quite quickly.

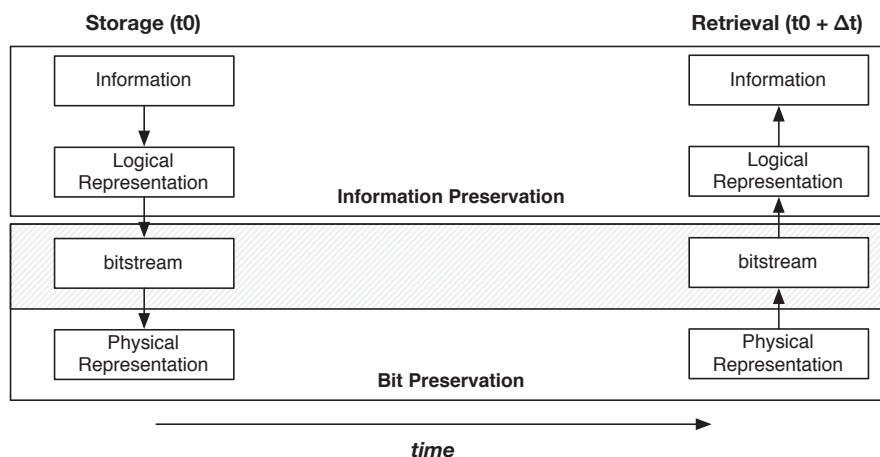


Figure 8: Level of concern of hardware obsolescence in the digital preservation communication model: preservation of bitstream accessibility (Image: F. Müller)

Similar to the case of material decay, no conclusive numbers can be provided for hardware obsolescence. However, the situation is easier to handle than that of material decay. An archive contains thousands (or more) of media that store data. The uncertainty about the state of decay of these media concerns every single one of them. But the archive will not have thousands of different media types – it will have only one or several storage technologies. The obsolescence of hardware is a problem only per technology. If the technologies employed in the archive are monitored regarding their availability on the market and their support in current computing platforms, the degree by which the archive is threatened by hardware obsolescence can be assessed reliably. Without proper monitoring, however, the archive may well encounter an unpleasant surprise.

### 2.1.3 Software Obsolescence

Software obsolescence is the most diffuse of the three problematic aspects, both in scope and interpretation. As stated before, digital documents consist of a bitstream. If the information contained in a digital document should be interpretable (i.e. used in an information process), the bitstream must be interpreted and the correct output must be produced. If the knowledge on how this interpretation works is lost, the bitstream will not yield information<sup>9</sup>. Software obsolescence concerns the information preservation layer (see Figure 9). There are considerable differences among the various document formats, especially when we regard their interpretation dependencies. Consider a text-only file, on one side, and the save-game of a computer game on the other side<sup>10</sup>. The text document can be viewed using any text editor. The only thing the text document contains is text in a specific encoding<sup>11</sup>. This is different for the save-game. Viewing it means continuing to play the game. The save-game all by itself contains no valuable information – it is only in conjunction with the computer game it was created from that it is useful. Whether a document depends strongly on a specific application affects the way in which it must be

---

<sup>9</sup>Rothenberg [30] has treated this aspect in detail. Since this topic has received broad attention, it is not discussed in detail here. For those unfamiliar with the related work, a simple analogy may be helpful. This text is a sequence of digits from the latin alphabet (and some additional symbols, which we do not consider in this example). You are able to interpret them, i.e. you have sufficient command of the English language to understand the text that has been written. If you were completely unfamiliar with English, but would be able to read the latin alphabet, this text would be unintelligible to you, since it would actually only be a sequence of digits for which you have no understanding. A computer without any software for interpreting documents is comparable to a person who can identify latin characters, but does not speak any language that uses it

<sup>10</sup>A save-game stores the progress of the player in a computer game and the state the game world is in at the time when the user creates the save-game. Save-games let the user resume the game where she left it at a later time.

<sup>11</sup>A text encoding governs which bit sequences represent which textual characters. The ASCII encoding is very minimal and can only encode 128 (7 bit ASCII) or 256 (8 bit ASCII) textual characters. The UNICODE encoding is larger and can encode up to 65.535 textual characters.

preserved. The stronger the dependency, the less likely it is that preservation of the mere document is sufficient.

It is undisputed that *in principle*, software obsolescence can be catastrophic. But the true scope of the problem has rather recently been put into perspective, e.g. by Rusbridge [46]. He doubts that software formats become obsolete quickly, and finds that the term *software format* is rather unclear. He proposes to distinguish several categories of formats, among them formats created by hardware devices (scanners etc.), formats from open source projects or based on standards (Open Document Format, PDF, TIFF, etc.), formats of proprietary consumer software, and file formats protected by encryption such as in the case of Digital Rights Management. Not all of these categories are in danger of obsolescence to the same degree. Open source projects have good prospects for providing long-lasting formats. Among the more endangered are the proprietary formats from consumer software. Encrypted documents are a special case, since they do not only depend on the continued availability of the cryptographic technique used for encryption, but also on the key required for decryption<sup>12</sup>. As with material decay and hardware obsolescence, awareness of the principal problem can greatly reduce its effective impact. If document formats are chosen carefully, their obsolescence is unlikely to be imminent. It will be shown later that developments in the field of emulation, which will be discussed as a proposed solution for software obsolescence, have already been very successful in maintaining interpretability of old formats.

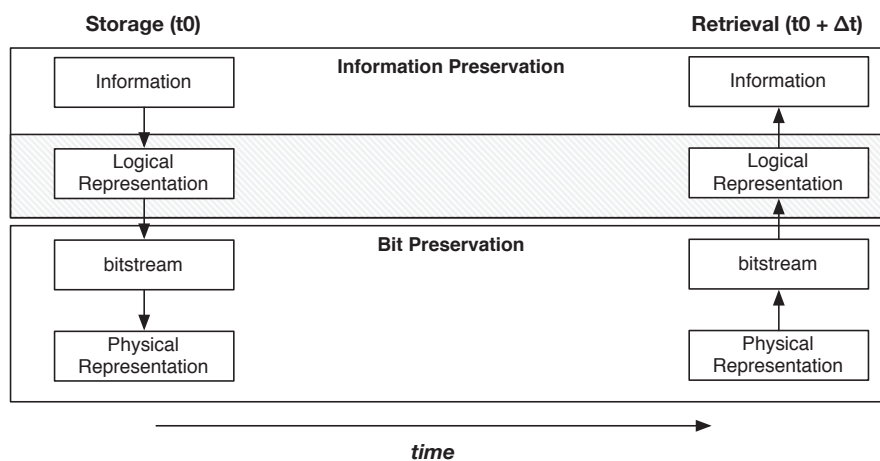


Figure 9: Level of concern of software obsolescence in the digital preservation communication model: logical representation (Image: F. Müller)

<sup>12</sup>It could be argued that this is not a case of software obsolescence, since an encrypted document is intended to not be interpretable without the proper key, and the loss of the key is not the equivalent of the obsolescence of a technology

In 2009, the Imaging and Media Lab has consulted an institution from the cultural heritage sector in finding a solution for organizing their digital photograph collection. The evaluation of several software candidates illustrates the importance of a careful choice regarding the employed software. In the final evaluation phase, 7 software systems were analyzed, four of them image databases, and three of them image organizers. Apart from considerations such as functional range, cost, suitability for the expected task, scalability, etc., several criteria were directly concerned with software obsolescence (or rather, with ensuring that should it occur, its impact would not be catastrophic). The following list of criteria that must be met by software candidates has been compiled in the course of the mandate. The organization in question is called the *organization*. The software candidate is called *software*, the complete image collection with all (existing and continuously created) metadata is called *information*.

- **Storage.** The information should ideally be stored in a manner that allows its access even without the software. A positive example would be the transparent storage of all information on the file system (e.g. images as image files, accompanying metadata as text-only or XML files). A negative example would be the storage of all information in a proprietary database.
- **Export.** The software must offer the ability to completely export the information in an openly accessible format. The organization must at all times have the possibility to use the entire information with a different software (which, in turn, must provide a suitable import mechanism).
- **Dependencies.** External dependencies – such as runtime-environments and database systems – should be minimal. If external dependencies exist, they should be as open and as standardized as possible. The reason is that the obsolescence of any dependency will cause the obsolescence of the software.
- **Sustainability.** The context in which the software is developed must be considered. One of the software candidates in the IML mandate was the development of an individual, and not open source. If the further development and support for a software depends on one single individual, the risk of obsolescence is considerably higher.
- **Licensing.** The software should not contain a time-dependent activation condition. If, for example, the software only works for one year, and further operation depends on the purchasing of a license for an additional year, the risk of lacking future funding or a development and support cease is considerable.

This is to emphasize again that software obsolescence can neither be predicted, nor completely excluded. However, there are many factors in control of

the archive that greatly influence the probability of obsolescence taking place, and the possible impact it has. Such a set of criteria is comparable to the various software format categories Rusbridge has proposed. Depending on which criteria the software fulfills, the software format it generates can be considered more or less probable to become obsolete.

#### 2.1.4 Further Problems

The three problems discussed so far – material decay, hardware and software obsolescence – have been called fundamental. They are especially time-dependent, and affect the core technical infrastructure of the archive. There are further problems for digital preservation that are different regarding these two characteristics. Nevertheless, they are very important for successful digital preservation. Considering all such problems would lead too far in the context of this thesis. Also, the selected problems will not be covered in-depth, with the exception of authenticity, which will be the focal point of section 2.5.

**Authenticity.** Merely preserving documents is not sufficient. In order to make use of the documents preserved, their authenticity must often be assessed. For example, how useful would a preserved contractual document be if we were unable to tell whether it has been modified since the signing of the contract or not? Ensuring document authenticity is challenging in digital archives, and the task requires different methods compared to the analogue archive. This problem will be treated in more detail in Section 2.5. It is not a problem of whether or not we can preserve information, but of how valuable this information can be.

**Document delimitation.** Digital documents can range from simple text documents to complex web pages. That a delimitation of documents is not always trivial is illustrated especially in the context of the web. When we see a web page in our browser, the core document we see is an HTML<sup>13</sup> document. In the early days of the web, web pages were just HTML files stored as such on the server, containing text with some images and links to other documents. They were comparable to conventional documents: these also contain text and images, and external links are provided as references, not as clickable links. If we look at a web page today, however, this is quite different. Take the example of the New York Times main page<sup>14</sup> (in newspaper terms, the front page). When we access it with a browser, we still obtain an HTML document. However, the manner in which this is generated cannot be described as static.

The web page in question contains headlines and links to articles, a list of the most recent brief news, a list of the most recent relevant blog entries, a ranking

---

<sup>13</sup>Hypertext Markup Language. HTML is a markup language that was developed from the Standardized General Markup Language SGML, which, by now, has largely been replaced by the Extensible Markup Language XML

<sup>14</sup>The New York Times, <http://www.nytimes.com>, accessed October 2011



of the articles that are currently most popular, advertisements, real-time stock information, and much more. Since the various components represent dynamic information, the web page – the document in question – is subject to constant change. This change depends on the time of the page load, the location from where it was loaded, the request from where it came, and possibly additional factors. In essence, every time someone accesses the New York Times webpage, an individual document is created and sent to the user. How, then, should one proceed in preserving the New York Times web page on any given specific day? In order to fully preserve it, one would probably have to preserve not the document that is sent to the users, but the server software that creates the documents. But then, what is to be preserved is no longer a document, but rather an information system (and, since it includes external information systems, such as several advertisement providers, the dependencies are numerous). And document dynamics are not limited to the server side – through the use of modern scripting technologies such as AJAX (Asynchronous Javascript and XML), the client obtains from the server not a dynamically generated document, but a dynamically generated (client-side) document generation system. Certainly, this does not make preservation of such web pages impossible, nor does it make the conception of digital documents impossible to maintain. What is emphasized, however, is that what we perceive as a single document is only an instance out of a set of many possible, similar documents. Whether the differences between them are significant is left to others to evaluate.

**Metadata.** Metadata itself is not a problem, but rather insufficient care of it. Metadata is data about data. It provides information about what the data is about, how it is structured, and how it is to be interpreted. Take the example of an MP3 (MPEG-1 Audio Layer 3) music file. In its simplest form, an MP3 file consists of a series of MP3 headers<sup>15</sup> and MP3 data blocks. The MP3 data blocks contain the (compressed) audio data, and each MP3 data block is preceded by an MP3 header, which supports the decoding algorithm with properly identifying the various audio data parts. Imagine you have an MP3 library on your computer consisting of 1000 MP3 files of this form, and imagine further that you know that you have Bach's Goldberg Variations in a wonderful interpretation by Glenn Gould, and would like to listen to it. How will you find the appropriate file?

Probably, you have carefully named all your files according to their content and organized them into folders and subfolders. You are able to locate the Goldberg recording by browsing your hard disk. In that case, the file names and the folders and subfolders into which the files are organized can be said to be the metadata about your music library – they help you to tell which file contains what. One could argue that it would be preferable if the metadata that you have

---

<sup>15</sup>In this context, the MP3 headers are not seen as metadata, but rather as an integral part of the audio data

were contained in the files themselves. There are several arguments for that, here, only one is presented. Suppose you share one of your files with someone else<sup>16</sup>. If you share just the file, the metadata contained in the folder structure will be lost. MP3 files support not only the storage of basic audio and control data, but also of metadata. The header and data sections can be preceded by a metadata section. This metadata section contains a set of relevant attributes, such as the track title, the artist<sup>17</sup>, the genre, etc. In short, all the information you need for retrieving the recording of your choice is contained in the MP3 files, and a music library management software will be able to provide you with a searchable database of your entire digital music collection based on the metadata contained in the files.

In the preservation of digital documents, metadata is crucial. In the quest for a coherent view of digital preservation, the development and standardization of metadata schemes have received a lot of attention. Metadata is not the focus of this thesis, although it must be emphasized that digital preservation without metadata is close to unthinkable. However, we consider the introduction of appropriate metadata schemes to concern the application of our digital archiving system. The technical and conceptual foundation provides the means necessary to do so, but no detailed implementation.

**Licensing.** Licensing is an important factor in digital preservation. At a first level, the ability to make copies of digital documents has a direct impact on the ability to preserve them: the more copies, the safer. Restrictive distribution licenses limit the number of copies that will be made of documents. But there is another level, and it is most important regarding software obsolescence. Consider a software  $S$  that creates some sort of document  $\mathbb{D}$  at a time  $t_0$ . At a future time, the following may be true for  $S$ :

1.  $S$  is available in its original form and runs on current computer platforms
2.  $S$  is available in its original form, but does not run on current computer platforms
3.  $S$  is available in a more recent version, runs on current computer platforms, and is compatible with  $\mathbb{D}$
4.  $S$  is available in a more recent version, runs on current computer platforms, and is partly compatible with  $\mathbb{D}$
5.  $S$  is available in a more recent version, runs on current computer platforms, but is incompatible with  $\mathbb{D}$
6.  $S$  is unavailable in its original form, no new version is available

---

<sup>16</sup>Of course, you would *legally* share it

<sup>17</sup>Would it be Bach or Gould? How appropriate a metadata schema is depends on your point of view

We could think of more cases, but the ones listed suffice for clarifying the meaning of licensing for software obsolescence – we focus on the cases where things go bad. In case 2, the problem would be that the program only exists in a binary form, and would have to be executed on an old computing platform. If the source code of the software were available, it probably could be recompiled using current computers. In case 5, licensing will not help much – if the developers of a software have decided to not include backwards compatibility, this may be a bad decision. Case 6 could happen if the software company goes out of business or discontinues the software entirely, and is not able or willing to provide copies of the old software. The thing with open source software is this: if a software is open source, it means the entire software in its human-readable, modifiable form can be acquired by anyone. If you have such a program, it means that you can recompile it for future platforms. Compilers and such have been shown to be much more persistent in time than the programs created with them. Second – and this is probably more important – open source software does not depend on specific developers, but rather on the development community as a whole. If any open source software is attractive enough, there will be enough people willing to develop it further, filling possibly opening gaps resulting from previous developers losing the interest or ability of development. The importance of using open source licensing schemes has been emphasized, among others, by Kuny and Rosenthal. Several successful digital preservation software systems – LOCKSS<sup>18</sup> from Stanford University, DSpace<sup>19</sup> from the MIT, XENA<sup>20</sup> of the National Archives of Australia, Fedora<sup>21</sup> from the DuraSpace organization – are developed as open source software. Licensing questions are relevant not only in digital preservation, but also in the wider field of digital sustainability.

**Economic failure.** Economic considerations are important in digital archiving. Digital preservation is a continuous activity that involves many organizational processes, staff, and infrastructure. Hence, it is an activity that produces continuous costs. If for any reasons, the financial backing of an archive is endangered or even fails for a period of time, the information in the archive is endangered. Note that this is not only valid for expected costs, but also for unexpected costs resulting from any of the problems of digital archiving introduced in this section. In a manner of speaking, the success or failure of digital preservation is determined not now, but in the future – if the information can be regained, digital preservation was successful, otherwise, it has failed. The fundamental problems in digital preservation are of a technical nature, but their solution in the future, should it be required, is not entirely technical. Suppose

---

<sup>18</sup>Lots of Copies Keep Stuff Safe, <http://lockss.stanford.edu>, accessed October 2011

<sup>19</sup>DSpace, <http://www.dspace.org>, accessed October 2011

<sup>20</sup>XML Electronic Normalising for Archives, <http://xena.sourceforge.net>, accessed October 2011

<sup>21</sup>Flexible Extensible Digital Object Repository Architecture, <http://www.fedora-commons.org>, accessed October 2011

that a storage technology has gone obsolete, that an archive has not taken appropriate countermeasures, and that at the time the information is to be accessed, no devices exist to read the media that are still intact. Suppose further that the knowledge that was required to engineer the technology is still available. Will the obsolescence of the technology make digital preservation fail in that case? Possibly, there will be a technically feasible way to regain the information. But whether this way will be economically affordable is another question. In short, financial considerations are not only crucial for the operation of the archive, but to a large extent also determine the impact of the other problems of digital preservation.

### **2.1.5 Summary**

Material decay, hardware obsolescence and software obsolescence are the fundamental technical problems of digital preservation. Together, they form what we call the problematic triangle (see Figure 3 at the beginning of this study). While they are fundamental and universally acknowledged, their true scope remains debated. It is especially hard to give estimates of exact time horizons within which one or several of them become problematic. Thus, these problems should be considered as follows. The five year hardware obsolescence timescale proposed by Rothenberg gives a good idea of the imminence of the problems of digital preservation. While it should not be taken literally (giving exact numbers is just not possible), it does suggest that when concerned with digital preservation, there are two imperatives: (1) the choice of a suitable storage system and of employed software formats must be made carefully and informed. Especially, it must include thoughts about an exit strategy, which will have to be followed should one of the corners of the problematic triangle become apparent. (2) The material state of the carriers, the hardware compatibility and the software format compatibility must be assessed periodically, such that possible problems are detected early and can be solved diligently.

## 2.2 The Permanent Medium Approach

Several strategies have emerged to cope with the fundamental problems of digital preservation. A rather complete account is given by Rosenthaler [47]: (1) hard copies, (2) standardization, (3) computer museums, (4) migration, (5) emulation and virtual machines, (6) permanent media and (7) digital archeology. Strategies 1 to 3 are not promising: relying on hard copies is the equivalent of not going digital, the hope that standardization will solve the underlying problems may well prove to be an illusion, and computer museums may delight enthusiasts, but cannot sincerely be considered as the way we handle our archives (although it should be noted that the preservation of operating obsolete machinery may prove useful in the future). The last proposition, digital archeology, i.e. the forensic reconstruction of data and information from obsolete and possibly damaged media, may be the last resort in some cases where preservation has gone bad, but will not scale sufficiently. Of the remaining strategies, migration and emulation are by far the most popular. For reasons of brevity, they are not discussed in detail at this point. Readers are referred to good accounts of the two strategies e.g. by Rothenberg [48] or more recently Borghoff [49] and Dobratz [50].

We will focus on the permanent medium approach, since Peviar follows this approach. Of all the approaches, it most directly challenges the fundamental problems of digital preservation. In order to make sure that digital preservation does not fail due to these problems, any permanent medium must make sure that (a) it is not subject to irrecoverable degradation, (b) it is not subject to hardware obsolescence, and (c) it is not subject to software obsolescence. In the case of Peviar, this is achieved by providing an ultra-stable carrier, by using a visual interface for data detection, and through the possibilities of a hybrid medium.

### 2.2.1 Ultra-stable carrier

The first prerequisite is a carrier that remains stable over very long time spans. What is meant by long time span depends on the context. In the OAIS reference model, a long time span was defined as a time span in which technological progress has a significant impact on the infrastructure of the archive. In the context of historical carriers of information from before the digital age, this does not compare. The most prominent (and successful) information carrier is paper. Prior to the nineteenth century, paper production used linen, hemp, cotton or wool as base materials. Wood only became the base material in the nineteenth century, when large-scale industrial production of paper was made possible through the introduction of several paper-production machines. The new production techniques made paper cheaper and more widely available. It was noted only in the twentieth century, however, that such paper would turn acidic over time, leading to a relatively fast deterioration of the material (fading, brittleness). New production techniques were developed for acid-free or alkalic

paper, usually through the introduction of an alkalic base (calcium carbonate, magnesium oxide). For existing acidic paper, de-acidification techniques were developed that allow to increase the lifetime of acidic paper considerably [51]. While old books clearly show signs of ageing, the time horizon of 100 years is easily achieved by them. Table 2 gives an overview of the book collection of the University of Basel library. From a total of a little more than three million books, almost 300.000 are more than 100 years old (historical collection), and 100.000 are more than 200 years old.

Total collection	3.100.000
Historical book collection	296.811
Before 1501 (incunables)	2.819
16th century	22.005
17th century	23.503
18th century	46.939
19th century	201.422

*Table 2:* Historical book collection of the University of Basel's University Library as of the year 2006. All books printed before 1900 are considered part of the historical collection. 123 titles could not be chronologically classified, but are considered part of the historical book collection (Source: [52])

The conception of long-term as a time period in which the impact of technological change is relevant to the infrastructure of the archive is not suitable for the digital preservation approach of a permanent medium. Its aim is the elimination of the impact of technological change, and hence it should not be measured against it. The time horizon suitable for Peviar is closer to the ones in the historic book collection, where long-term is in the order of centuries, not decades. For the purpose of Peviar, we aim at long-term preservation for at least 100 years. This is also the time horizon that was relevant in a Peviar application, as for example indicated by the Loveletters project briefly introduced in section 2.4. Consider that this is far more time than has passed since the invention of the first electronic digital computer, and over three times the span since the introduction of the personal computer.

### 2.2.2 Visual Interface

Digital data is always physically represented as analogue signals, and we call the nature of the physical properties used for representation the domain of representation. Magnetic media use the magnetic domain, i.e. the physical property of magnetization. Optical media use the optical domain, i.e. the physical properties associated with electromagnetic radiation in or near the visible range. In contrast to the magnetic domain, where the physical property to represent data is always magnetization, the optical domain offers several properties that can be

used to represent data: the radiation wavelength (light color), the radiation amplitude (light intensity), the radiation phase (spatial origin of the light) and light polarization. Depending on which of the properties are used, we can distinguish several subgroups in the optical domain. Holographic recording, for example, stores all electromagnetic radiation properties (amplitude, wavelength, phase and polarization). CD-ROM/-RWs and DVD-ROM/-RWs can differ in what properties they use: some use only amplitude (based on presence or absence of a metallic or dye layer, for example), some rewritable media use polarization or even magneto-optical properties.

What is required for our approach is not only an optical, but a visual medium. Visual media can be viewed as a subgroup of optical media. This classification does not happen in exclusion of the proposed classification according to the electromagnetic properties used for data representation, but rather in addition to it. For an optical medium to also be a visual medium, the data representation must be *visible* to the human observer, and the totality of the data representation properties must form an *image*, i.e. the medium appears to the human observer to carry one single image. This is different for other optical media, e.g. CD-ROMs. When looking at a CD, the observer simply sees a shiny surface, the physical properties used to represent data are not visible. When looking at a visual carrier, the observer immediately sees an image (whether it makes sense to her or not). As will be shown in the following sections, the advantages of this are twofold: it allows the inclusion of directly perceivable information, and it provides the base for simple data access via image capture.

Storage technologies consist of a carrier (the medium) and a device that can record and read it. For concise reading, we call the device for recording and reading the *drive*. By applying specific physical properties to a carrier, digital data is materialized more or less permanently. Upon read-back of the data, the drive detects the physical properties applied earlier and reconstructs the original data, ultimately transferring it to the device responsible for further processing (interpretation, rendering). The problem of hardware obsolescence concerns the continued availability (and compatibility) of a given storage technology, i.e. of a carrier and its drive, as has been described in section 2.1.2. In our Peviar scenario, the future absence of current drives is even assumed.

How this future absence of current drives affects possible read-back is the separating feature between visual and non-visual media. Non-visual media presuppose the availability of specific drives (and with them specific drive firmware, specific drive connectors and operating system drivers). If we want to read a CD-ROM, we need a CD drive containing a small laser that is able to properly detect the presence or absence of engravings on the disk. Building such a drive according to the specification of the CD-ROM technology from scratch is expensive and difficult. This is very different for a visual carrier, for the only technology it presupposes are: (1) the ability to capture images of the real world in sufficient resolution, and (2) the ability to process these captured images with programmable software. Once the visual carrier has been captured by a camera

or scanner, the data can be reconstructed from the resulting (digital) image using only image processing software (see the next section for possibilities of ensuring that this software will be available in the future). We can see that the presupposed abilities - image capture and processing - are rather general. In fact, they are technological families or domains rather than specific technologies. We consider the assumption that these will be available safe. First, capturing images of the real world has been an object of desire for mankind since a long time<sup>22</sup>, and following its realization through the invention of photography 150 years ago, it has made enormous progress. There is no apparent reason why this ability should not continue to be considered very important. Second, software processing of captured images requires only the availability of computers, i.e. machines that are universally programmable to calculate the things that can be calculated.

### 2.2.3 Hybrid Medium

A visual medium is apparent as an image, and that image can depict anything images can depict. We are entirely free in designing the internal structure of the image, and this has an important consequence: we can conceive a hybrid medium. The term hybrid, in this context, is used to refer to a medium that contains both analogue and digital information.

In traditional visual media, there is only analogue information: images and written text<sup>23</sup>. If we want to use a visual medium for digital information, all we must do is use the visual properties to represent data. A proven method for representing data in the visual domain are barcodes. First operational barcode systems were developed by Silver and Woodland around 1950 [53] and have become ubiquitous in the labeling and organization of objects (especially in the context of commodities, such as in supermarkets). Conventional barcodes are one-dimensional (they *encode information in one dimension*, see Figure 10 (a)), which is sufficient for encoding short numbers (and, in some variants, several characters). But if more information is to be encoded, both dimensions available on visual media must be used, and hence, two-dimensional barcodes such as the QR Code<sup>24</sup> have been introduced, which encode information in two dimensions (see Figure 10 (b)).

---

<sup>22</sup>The first camera obscura was built around the year 1000 CE. It was able to project an image onto a plane, but not to make it permanent.

<sup>23</sup>The case of text is somewhat tricky. Text is digital, but in a sense, it also is analogue, since decoding of written text actually happens in our brain. Colloquially speaking, the image of the text travels to our perceptive apparatus as an analogue image, and is then analyzed in terms of its digital contents. In the sense that written text is directly perceptible for humans (given they have been trained properly), it is analogue information.

<sup>24</sup>QuickResponse Code, developed by Denso-Wave in 1994, later standardized by ISO [54]





Figure 10: Comparison of one- and two-dimensional barcodes. The one-dimensional Universal Product Code (UPC-B) in subfigure (a) encodes information in the horizontal dimension, the two-dimensional QR-code in subfigure (b) encodes information both in the horizontal and vertical dimension (Image: F. Müller)

Apart from being optimized for machine recognition, 2D barcodes have a high storage density compared to written text. Written text is always represented using a font. Each font is a set of symbols than can be described as a certain configuration of an  $n$  by  $m$  matrix. In Figure 11 (a), the characters of a VGA font are listed. There are a total of 256 characters (this is the case for any VGA font). Figure 11 (b) shows an 8 by 8 matrix used to display a single font character<sup>25</sup>. The matrix is sufficient for encoding the entire VGA font character set. Figure 11 (c) shows the use of a 3 by 3 matrix for a 2D barcode. If 8 cells are used to encode a character and one cell is used to encode parity bit information, the 3 by 3 matrix is able to encode the entire VGA font character set, as well. Compared to the 8 by 8 font matrix, it uses only 10 percent of the space. In the investigation of photographic material as an information carrier, Altman et. al. measure a higher information density for information encoded in the form of a barcode compared to textual encoding [55].

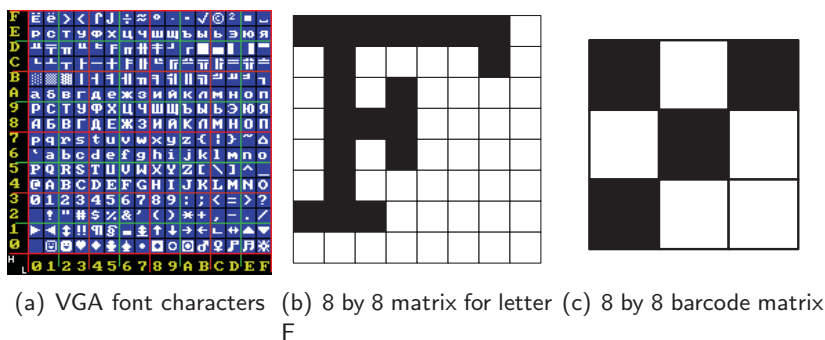


Figure 11: VGA fonts

<sup>25</sup>VGA fonts always use a matrix of width 8 or 9, the height varies up to 32.

By using arbitrary images and two-dimensional barcodes, a hybrid medium allows the combination of analogue and digital information. In a digital archiving system, the emphasis lies on the digital – what the system should actually do is preserve the digital information. The analogue information is not placed for its proper preservation, but included to provide documentation and ensure continued access to the digital information. Encoding information as digital data is conducted using rules for representation. The interpretation of digital data that was encoded using such rules is not self-evident. If we do not know what the bitstream actually *means* (i.e. according to which rules it was derived from the original information), we are unable to make use of it. This is where the accompanying analogue information comes in. In a manner directly intelligible, it documents the inner workings of the digital code. A very simple example is provided in Figure 12.

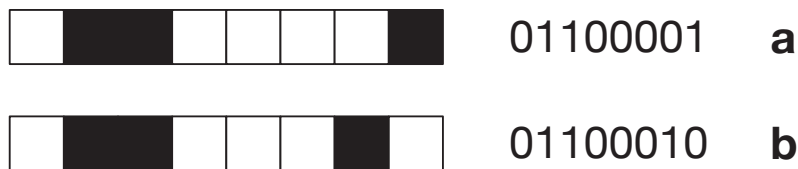


Figure 12: Self-documentation of a hybrid medium. Analogue information (and common sense) are used to explain the digital encoding of textual information ('a' is digitally encoded as '01100001', and represented in the visual domain as a series of squares with the colors 'wbbwwwwb' (w=white, b=black) (Image: F. Müller)

The example demonstrates how a hybrid medium can document itself, or rather, the specifics of the digital information contained on it. While the example is very simple in several regards, it does illustrate the principle possibility of self-documentation. An important aspect in that regard is self-intelligibility. In order to provide documentation about something, the manner in which it is provided must be intelligible without further ado. If this was not the case, the documentation itself would not be understood. A very clear example case of this problem has received prominence in the context of the human exploration of space. In their 1972 and 1973 Pioneer 11 and 12 missions, the U.S. National Aeronautics and Space Administration (NASA) attached the so-called *Pioneer plaque* to their spacecrafts, a 9 by 6 inch (22.86 by 15.24 centimeters) golden plate schematically depicted in Figure 13. The purpose of the plate was to communicate to “scientifically educated inhabitants of some other star system”<sup>26</sup> details about from where, when and by whom the Pioneer spacecraft were launched. In a manner of speaking, it is a very sophisticated cave

<sup>26</sup>See NASA pioneer mission archive, <http://www.nasa.gov/centers/ames/missions/archive/pioneer.html>, accessed October 2011

painting<sup>27</sup>.

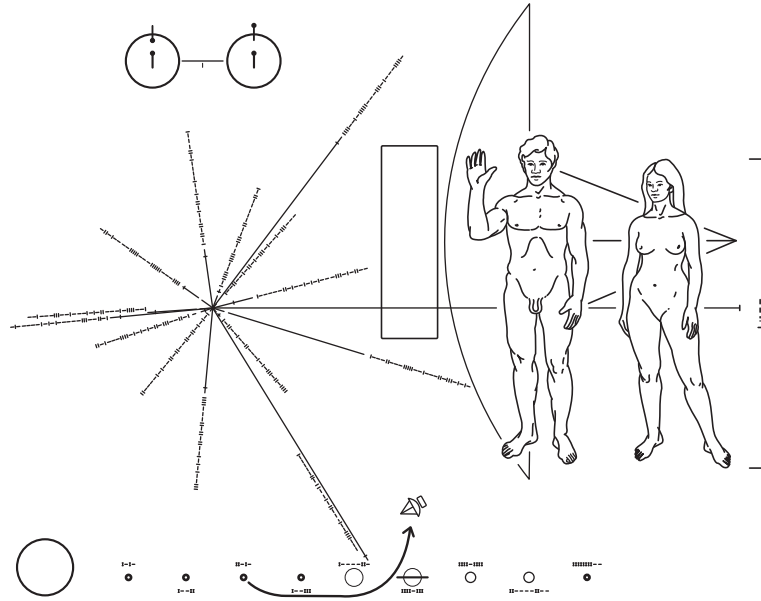


Figure 13: The Pioneer plaque (Image: NASA)

There have been attempts to provide a formal model for the concept illustrated by the Pioneer plaque. Cheney et al. [56] have investigated the use of universal algebra as a means to describe the dependencies of digital objects. Tzitzikas et al. [57] have introduced a formal model for intelligibility. It is useful for transparently modelling dependencies in a digital preservation environment and denotes a lack in dependency understanding as an *intelligibility gap*. In other words: an intelligibility gap occurs when something (e.g. the documentation) is not understood without additional knowledge. Self-documentation must be self-intelligible.

What we consider to be self-intelligible depends mostly on what we assume about our communication partner. If we assume nothing at all, then we are unable to communicate. In the case of the Pioneer plaque, which is destined to hypothetical extra-terrestrial intelligence, the assumptions include that the recipients have what we consider to be advanced scientific knowledge about the world (e.g. hyperfine transition of hydrogen), that they are able to interpret the plaque as a whole and solve its riddle, that they understand the arrow symbol, etc. In order to come up with a reliably self-intelligible documentation in the context of digital preservation, we must discuss which assumptions regarding

<sup>27</sup>The pioneer plaque has been shown to several scientists on Earth. None of them have understood the message completely <http://www.strange-loops.com/scitalktoaliens.html>, accessed October 2011

future information recipients are valid, and which are not.

Since Peviar is not developed to send information into outer space, we assume that the recipients will be our future generations. The self-documentation of Peviar can vary – it can be formal or informal, extensive or concise, in English or in German. In the *Loveletters to the Future* (short *Loveletters*) project, a set for self-documentation has been created [58]. The most important assumption was that future recipients will have command of the English language. Languages have a long tradition of being a suitable encoding for information. Today, languages like ancient Greek or Latin are still understood by some and the object of scholarly research. It is possible to decipher lost languages a posteriori, even if sometimes only with the help of an empirical translation base, as in the case of the Rosetta stone.

## 2.3 Color Microfilm

The properties of a permanent medium discussed so far are found in several photographic materials, most notably, in color microfilm. Compared to other materials offering permanence, a visual interface and a hybrid character (e.g. paper), it is superior in terms of stability and achievable data density.

The term *microfilm* (sometimes *microform*) was introduced for specific applications of photographic materials, namely the scaled-down reproduction (microproduction) of mostly textual documents on photographic film for preservation and transport. First experiments with microproduction of texts date back to 1839 [59]. Widespread applications were not developed until well into the twentieth century, when Eastman Kodak bought the first patented microfilm machine (Checkograph) from pioneer George McCarthy and started commercial operations with its *Recordac Division* in 1928. Early applications were the filming of newspapers, among them the Harvard University Library's *Foreign Newspaper Project*, which microfilmed several non-US newspapers on a daily basis and still operates today<sup>28</sup>. Microfilm was attractive because compared to paper, photographic materials offered a much better lifetime. In the second world war, microfilm was also used for other reasons. Documents could be sent over great distances at a fraction of their original size, making the transportation less expensive and less likely to be discovered. After the war, microfilm was not only widely used in libraries and archives because of its long stability and high information density (from on the 1950's until today). In a time where documents could not be distributed electronically at a large scale, microfilm was a very compact and cost-effective distribution medium, mostly in the form of blue diazo prints. In the 1970's, Computer Output Microfilm (COM) was introduced. It allowed the use of microfilm as an arbitrary printing device for computer-generated documents.

While microfilm was initially named after a specific application of photographic materials, it later became a synonym for photographic material that is especially well-suited for microproductions. Most importantly, microfilm has a high resolving power (it must be able to reproduce fine details at a very small scale). It therefore has a low granularity and is not very light-sensitive. It is usually thin compared to other films to save storage space. Also, it provides strong contrast adapted to the primary use for textual documents, which require a high-contrast reproduction due to their inherently poor contrast. Traditionally, microfilm was monochromatic. In 1982, Ilford (then owned by Ciba) introduced their Color Micrographic Film, called Cibachrome. In 1989, when Ilford was sold to International Papers, the product was renamed to Ilfochrome (Ilfochrome CMM). Kodak had been developing a color microfilm on the basis of its chromogenic Kodachrome process. However, the film did not reach the sharpness of monochromatic microfilm, and was not lightfast. The ability to create microproductions in color was important especially for cultural heritage

---

<sup>28</sup><http://www.newspaperarchive.com/>, accessed October 2011

applications. In the context of this work, we use *microfilm* to refer to the Ilford Color Micrographic Film[60] that we have used throughout the project.

Ilford produces two variants of the film, a type M (Master) variant with high contrast for the reproduction of the originals, and a type P (Print) variant with lower contrast for creating contact copies of master film for distribution purposes. The film has multiple layers containing stable Azo-dyes and is processed in the direct-positive silver-dye bleach process<sup>29</sup>. Once the film is developed, its lifetime essentially depends on the stability of the dyes. Over time, these dyes inevitably deteriorate. How fast this happens is determined by the storage conditions, most notably by (a) the presence or absence of light, (b) the temperature  $T$ , and (c) the relative humidity RH. The more light is present, the warmer and more humid it is, the faster the microfilm will degrade. When the dyes deteriorate, their intensity is reduced, and the image fades. We distinguish between *light fading*, fading caused by exposure to light, and *dark fading*, fading caused by temperature and humidity. In the context of preservation, dark fading is the more relevant property, since it can be assumed that archival materials can be kept in the dark.

The stability of materials over time is most precisely determined by natural ageing experiments. In such experiments, a sample base large enough is observed during its natural ageing process. For practical reasons, this method is limited to materials with rather short expected lifetime. If a material is expected to be stable for several decades or even centuries, natural ageing experiments are not feasible. Since microfilm is stable over long time periods, another setup for determining its expected lifetime must be employed. A common method is that of accelerated aging. In accelerated ageing experiments (AAE), the material under investigation is kept under conditions that speed up its deterioration (i.e. elevated temperature and RH in the case of microfilm) and make the deterioration process observable within a period of several months. A deterioration threshold is determined beyond which the film is considered unusable. The expected lifespan is the time it takes the film to reach the threshold.

Such AAE do not measure the deterioration under normal archival storage conditions. In order to extrapolate from the high-stress conditions of the AAE to other environments, the Arrhenius and the Eyring model can be used. The base of the Arrhenius model, the Arrhenius equation, relates the rate constant of chemical reactions (the speed at which chemical reactions of a certain order occur) to the environment temperature. Given the deterioration (reaction speed) at a certain temperature, the deterioration at other temperatures can be calculated. The Eyring model is able to incorporate more factors than just temperature, e.g. the relative humidity. Given such a model, the natural ageing of microfilm can be inferred from the observed accelerated ageing. It should be noted, however, that such a setup depends on several assumptions [61]. Most

---

<sup>29</sup>The incorporated dyes are the main reason for the high resolving power, as they suppress the light scattering at the silver halide crystals.

importantly, it is assumed that the stress conditions causing the deterioration in the AAE setup are the same that cause deterioration during natural ageing. Consider the influence of bacteria or fungus. In natural ageing (especially given sufficient RH), they may cause damage to film materials, while in AAE, the high temperatures would kill bacteria and funguses. This does not in principle mean that the use of such models is inappropriate. Rather, it shows that the accuracy of predictions based on the model is limited, as the AAE setup is not able to capture natural ageing to the full extent.

Aware of the limitations, one can observe dye degradation at various high temperature and RH, and then extrapolate dye degradation at low temperatures and RH. The assumptions allowing the application of the mentioned extrapolation models are not undisputed for the case of color microfilm [62]. However, if the extrapolated expected lifetime is not interpreted too narrowly, this lifetime estimation method is as well founded as possible. It has been applied in several studies, most notably by Meyer [63] and Wilhelm [64]. Wilhelm gives lifetime estimates for various photographic materials. The threshold to determine the end of life is a 20% degradation of dye<sup>30</sup>. Wilhelm concludes that the expected lifetime of all Ilford microfilm types is more than 500 years.

This lifetime is clearly superior to the numbers discussed in the previous section about material decay. The ultra-high stability of the carrier is an important basis for creating a migration-free archiving system. Of course, the lifetime of microfilm is not eternal, but only very high. Thus, speaking of a migration-free system is not correct. We call Peviar a system that is virtually migration-free. In fact, after a period of several hundred years, the data would have to be copied to a new film (or other carrier). Given the current migration cycles of around 5 years, this difference is highly significant.

### 2.3.1 Film as an Information Carrier

The use of photographic materials in microfilming has first taken advantage of the fact that large quantities of information can be stored on film at a fraction of the physical size of their paper equivalent. The developments in microfilming ever since its commercial introduction in the 1920's have made it a valuable tool for efficiently preserving documents. In the early stages of computing (1950's, 1960's), photographic film was also considered as a means for recording digital data. During the development of their System/360 mainframe computer family, IBM formed a task group to provide an analysis of the various storage technologies currently available. However, they did not come to conclusive results, and hence the storage devices initially available for the System/360 at introduction time in 1964 included several form factor tapes, disks, disk packs, drums, and others. While these were all magnetic media, IBM also pursued the development of optical technologies [65].

---

<sup>30</sup>Actually, 20% of the least stable dye. The different dyes (cyan, magenta, yellow) have different stabilities.

In 1961, IBM demonstrated the *Walnut* system, which had been developed for the Central Intelligence Agency. Walnut was based on photographic film<sup>31</sup> and allowed the storage of almost 100 million photographs. The novelty was that it also allowed automated computer-based search and retrieval of these documents – an index of the entire contents of Walnut, along with a summary for every photograph (which of course had to be created manually), was stored in digital form.

Walnut was only able to store analogue documents, but soon after its demonstration, IBM created a new project branch for developing a digital storage based on photographic material ([65], p. 281). In 1967, the first machine of this new project, called the *IBM 1360 Photo Digital Storage and Retrieval System* (Photostore), was installed. It was capable of storing 1 terabit (approximately 125 gigabytes) of information – a number introduced some years earlier as a desideratum for performing simulations of nuclear explosions by the Atomic Energy Commission, for which IBM was developing the Photostore ([65], p. 281). At the time of its introduction, the Photostore did provide an unrivaled data capacity. Because it was a write-once system (developed film is unalterable, and files stored on film can only be read or destroyed - not altered) and because magnetic recording technologies (which are read-write) made rapid progress, it was not developed further. By the 1970's, the idea to use film for digital storage had become unpopular.

Aided by the development of two-dimensional (2D) barcodes, new applications for information storage on film have been created more recently. In the cinema industry, 2D barcodes are used to encode the audio track on the film in digital form (Dolby Digital [66] and Sony Dynamic Digital Sound [67]). IBM has patented a method for digitally encoding captured image metadata (EXIF) as a 2D barcode alongside the analogue exposure of images on film [68]. The idea to employ photographic microfilm specifically in digital preservation is even more recent. It was introduced [69] at a time when the problems of digital technology regarding the long-term preservation of information had been acknowledged and the term of the digital dark ages had been coined. Ever since, several research groups have been working on digital archiving systems based on a microfilm carrier, and first implementation projects have been concluded, such as preserving audiovisual cultural heritage documents [70]. Apart from the research at the Imaging and Media Lab, a microfilm-based digital preservation system is also under development at the Department for Signal Processing at the Braunschweig Technical University [71].

---

<sup>31</sup>It used the recently discovered Kalvar material, which is sensitive to UV-light and is developed by simply heating it up. Later, Kalvar was no longer used due to its poor long-term stability.



### 2.3.2 Information Capacity of Photographic Materials

Parallel to the development of photo-optic storage systems, film was investigated regarding its *information capacity*. An important tool in that regard was the Mathematical Theory of Communication (MTC) developed by Shannon, published in 1948 [72]. The MTC provided a quantitative approach to information and laid an important foundation for the entire field of information science. Among the important concepts introduced by Shannon are his (quantitative) *measure of information* and his description of a *general communication system*.

Information is measured in *bits*, short for *binary digit*, and is attributed to a message (so a message, or its parts, have an information amount). In the context of a communication situation, where a message travels from a sender to a receiver, the amount of information that the message carries is derived from the probability of the message being the message that it is:

$$I(m) = \log\left(\frac{1}{p(m)}\right) = -\log(p(m)) \quad (1)$$

where  $p(m)$  is the probability of the message being  $m$ , given that it could have been any message in the set of possible message  $M$ , and  $\log$  is the logarithm to the base of two (this would be different if we were looking at another than a binary measure of information). In other words, the amount of information a message carries depends on how *unexpected* it is. Consider a coin with two heads. Whenever it is tossed, the side facing up will display heads, so the probability of heads being on display is 1, and hence we have  $I(m) = -\log(1) = 0$ . Because the result is always fully expected, tossing a coin with two heads yields zero information. Consider a regular coin with heads and tails. If it is unbiased, the probability that after a toss, heads is facing up is  $\frac{1}{2}$ , and of tails facing up is equally  $\frac{1}{2}$ . The amount of information produced by tossing such a coin is  $I(m) = -\log(\frac{1}{2}) = 1$ , one bit. As the probability of a certain message being transmitted decreases, the information amount of that message increases. For example, an unbiased die will equiprobably produce a number of  $\{1, 2, 3, 4, 5, 6\}$ . The information gained by tossing it is always  $I(m) = -\log(\frac{1}{6}) = 2.58 \text{ bit}$ .

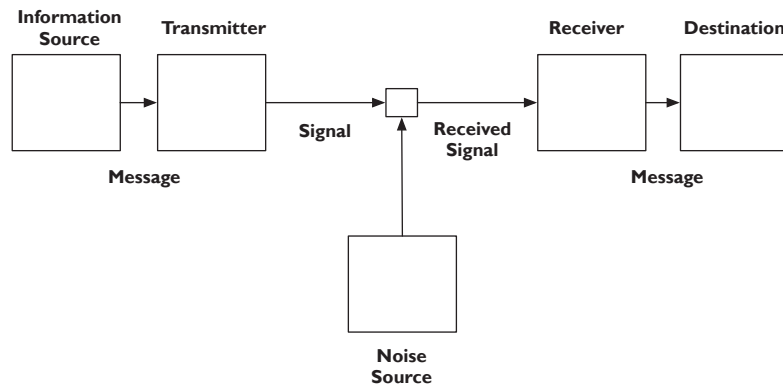


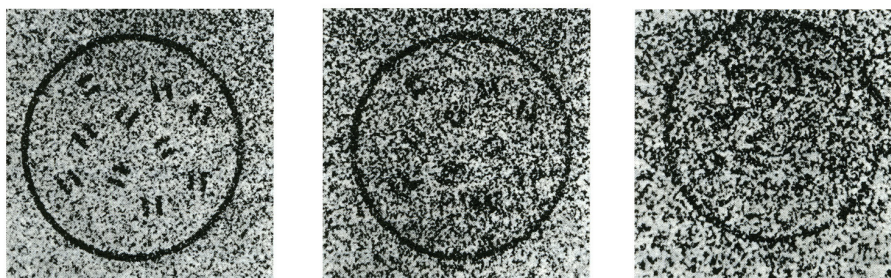
Figure 14: General Communication System as described by Shannon (Image: F. Müller)

The general communication system proposed by Shannon is depicted in figure 14. The information source is the producer of the message, i.e. it expresses information in the form of a message. The message is then given to the transmitter, which knows how to transmit it over the channel. The transmitter actually encodes the message in the form of a signal, such as an electric current. The signal travels over the channel, which is the (physical) connection between the transmitter and the receiver. The channel may be (and in reality always is) subject to noise from a noise source. At the receivers end of the channel, the signal is received, and the received signal is the original signal that has been subjected to the noise of the channel. The receiver decodes the received signal into a message, typically, this is the inverse operation of the transmitter. Once the receiver has created the message from the received signal, it is handed to the destination. The general communication system provides a model for all sorts of specific communication systems. Consider the telephone system, for example. Supposing that a connection between two telephones has been established, the person at one end of the communication (information source) speaks into the speaker, which will modulate the electric current of the telephone line according to the measured acoustic pressure (transmitter). The current modulation (signal) travels through the telephone line and is subject to the noise of the telephone line. At the other side of the communication, the telephone detects the modulated current as subjected to the noise (received signal) and demodulates it, creating acoustic pressure (message). The person on the other side (destination) then hears what the information source has said.

Based on these fundamental concepts, Shannon has formulated a theory for determining the capacity of a communication channel. The channel capacity is the amount of information that can be transmitted over a given channel, i.e. the number of symbols that can be successfully decoded without error within a certain time period. If we want to express it in bits, it depends on the rate of symbol transmission and the information amount per symbol in bits. The rate

of symbol transmission depends on the frequency response, and the amount of information per symbol is limited by the signal to noise ratio. Soon after publication, Shannon's concepts were applied to film in photographic research. By viewing photographic materials as a communication channel, it was possible to determine the relevant properties of film as an information recording material, most notably, its information capacity. It was stated that the channel capacity depends on the number of symbols transmissible in a certain time frame. This is different for film, where the capacity depends on the number of symbols transmissible in a certain spatial plane. The equivalents of frequency response and noise are the modulation transfer and the granularity of the photographic material.

Photographic image generation depends on the formation of silver from the silver halide present in photographic emulsions. If the silver halide has come into contact with light, it will turn into silver. In the development process, the silver is used in one or another way to form the image. In the case of black and white film, the silver remains in the film, while the silver halides are washed out – the result is a negative black and white image (the film is black where it had contact with light). In the case of color film, either the exposed silver halide is used to react with color couplers (as a result, dye is formed in the respective layer – chromogenic procedure), or the exposed silver halide is developed into silver and later used to wash out a proportional amount of dye in the respective layer (chromolytic procedure). Now, the silver halides are not distributed uniformly in the emulsion. This means that given two distinct areas of the emulsion of equal size that have been exposed to the same quantity of light, the amount of silver halide exposed to them will not be exactly the same. The greater the area and the smaller the silver halide grains, the smaller the statistical variance. Figure 15 illustrates the effect of granularity on sharpness. In essence, given a uniformly exposed area  $A$  of a certain film, the measured optical density after development may vary.



*Figure 15:* The effect of granularity on sharpness. Increasing granularity from left to right makes it harder to distinguish the feature markings from the background noise (Image: P. Glafkidès, [73], volume 1, p. 311)

The modulation transfer gives a measure of the fineness of detail that can

be reproduced with a certain film. If a perfectly sharp edge is imaged onto a film, its reproduction will never be as sharp as the original. Figure 16 shows the capture of a black-and-white edge exposed onto color microfilm. The blur visible at the edge is the result of the limited modulation transfer capability of the imaging process. Modulation transfer can be expressed in terms of *spatial frequency*. If a film is able to reproduce very sharp edges, it provides good modulation transfer at high spatial frequency. A related measure often specified for films is the number of line pairs per millimeter,  $lp/mm$ . This number is determined by placing series of lines on the film at increasing spatial frequency, i.e. with decreasing line widths and spacings in between. The number of line pairs per millimeter still visually distinguishable gives  $lp/mm$ .

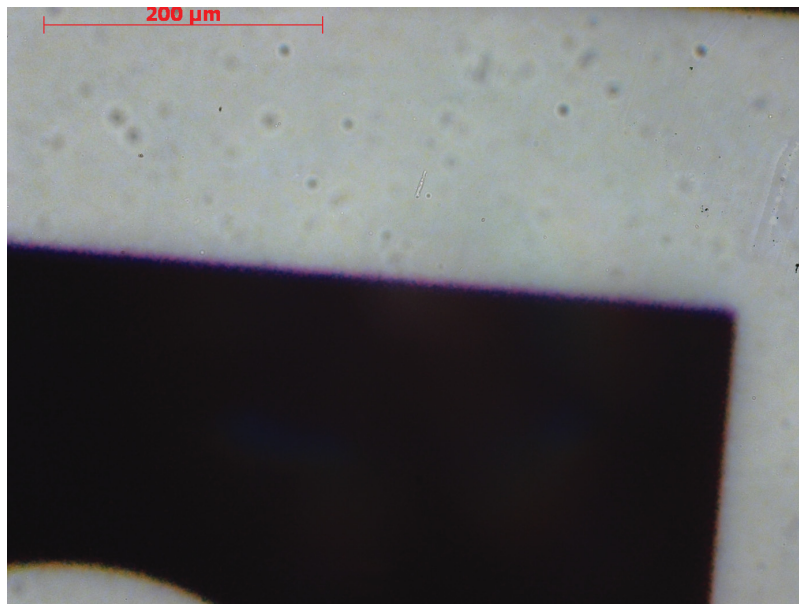


Figure 16: Limited modulation transfer illustrated at a black-white transition (Image: F. Müller)

We now have the two relevant properties of film as a communication channel, modulation transfer and granularity. The manner in which they determine the information capacity is as follows. The modulation transfer tells us the smallest structure size that is reliably reproducible (and detectable) on the film. In the context of information theory, we could say that it determines the symbol rate (in the spatial domain). The granularity, on the other hand, has an influence on the information amount each symbol can encode. In the binary case, a symbol is either a zero or a one, or, in film terms, a black spot or a white spot. But the film is able to reproduce not only black or white, but shades of gray<sup>32</sup>. In principle, the spots used to represent data could vary not only

<sup>32</sup>Color film consists of three layers, each of which can be treated as a black-and-white film.

between two levels (encoding  $\log(2) = 1$  bit), but between several gray levels (encoding  $\log(4) = 2$  bit at 4 gray levels,  $\log(16) = 4$  bit at 16 gray levels etc.). This is where the granularity has an effect. Due to the granularity, the optical densities resulting from uniform exposure vary. The smaller the area measured, the greater this variance is. At the same time, the more shades of gray are used, the smaller the differences in optical density between the signaling level. If the signaling levels and the granular variance are too close together, they become indistinguishable, and the signal to noise ratio is insufficient. Altman has come up with a simple heuristic model to determine the channel capacity of photographic film incorporating these two properties [55]. The number of signaling levels depends on the size of the spot as follows:

$$M = \frac{DS}{2k\bar{G}}\sqrt{A} + 1 \quad (2)$$

where  $DS$  is the *density scale*, i.e. the interval between the minimum and maximum density,  $\bar{G}$  is a constant for the average granularity,  $A$  is the aperture area of the scan, and  $k$  is the number of standard deviations by which the different levels are separated. The information that can be stored on a given area is then defined as:

$$I = N \log_2 M \quad (3)$$

where  $N$  is the number of spots. Combining the two above equations, we obtain:

$$I_{approx} = \frac{1}{2A} \log_2 A + constant \quad (4)$$

Altman concludes that the optimum packing density is reached when using binary recording at minimum cell size. Others come to similar conclusions, see the work of Shaw [74] or Glafkides [73].

In more recent approaches to film as an information carrier, the entire processing chain was analyzed. Early research has focused on the properties of photographic film exclusively. But when using film as an information carrier, the information must somehow be applied to the film (the film must be exposed or recorded), and at a later time, it must be read back. Amir [75] and Voges [71] propose channel models that include the recording component, so they treat the film and the recorder as one system, and their channel model serves to provide a measure of the capacity of that one system (and not film exclusively).

## 2.4 The PEVIAR Implementation

This section describes the implementation of Peviar using color microfilm. Figure 17 shows a Peviar microfilm produced for the *Loveletters to the Future* project<sup>33</sup>. The standard microfiche (105 × 148 mm, microfilm is also available in other physical formats) is divided into a header section (containing logos and ordering information) and a body section. The body section holds a 6 by 7 grid of image cells, each containing either analogue or digital information. For the Loveletters project, 14 videos were stored on microfilm. The fiche depicted stores one film entirely and one in part. The first cell on the first line contains the project logo. The second cell on the first line contains an analogue image that provides information about the following film, including technical meta-data. From the third cell of the first line to the fourth cell of the third line, a video sequence is encoded as 14 2D barcodes. The next film starts with the analogue meta-information in the fifth cell of the third line and spans to the next fiche not depicted here.

---

<sup>33</sup>In the run-up to the 2009 climate summit in Copenhagen, the PEVIAR technology was used in the *Loveletters to the Future* (LLTF) project. In a world-wide collaborative effort, internet users were asked to record messages to future generations, namely, the people living on earth in the year 2109. The messages should be conceived as love letters to the future, transmitting their readers the best wishes and thoughts about the now-current state of our planet. A jury picked approximately one hundred messages from all the submissions. These were to be sent to the future generations. Among them were 14 video messages and almost 100 letters with accompanying images. The mode of transmission was given. The messages were to be sealed in a time capsule, a custom-built container that provides space for several material objects. Among them would be the media that contain the messages to the future in the form of digital documents [58].

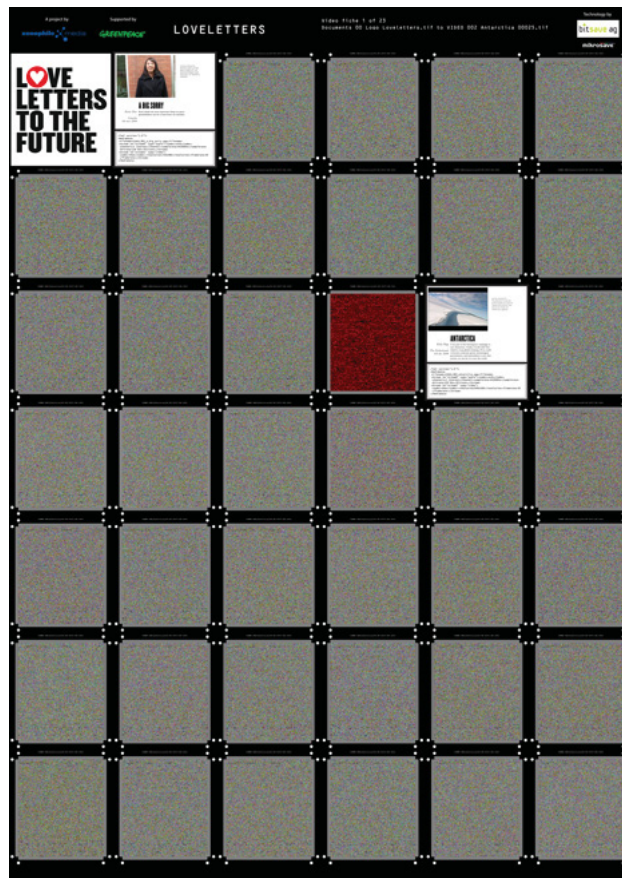


Figure 17: Peviar microfilm example as produced for the *Loveletters to the Future* project. It illustrates the hybrid nature of Peviar, containing both analogue (images, text) and digital (two-dimensional barcodes) information (Image: F. Müller)

The Peviar workflow, i.e. the production of microfilm storing digital information, is described in detail in Section 2.4.5. Here, a brief overview is given. Figure 18 provides a simplified illustration. The data that is to be stored with Peviar is first encoded as a sequence of 2D barcodes by the *Peviar encoder*. The result is a set of raster image files each containing a barcode. Any analogue information that should be included on the carrier must be prepared in the form of raster image files. Once these images are ready, they must be composed into one single image (as illustrated by the figure). This single image is then exposed to the microfilm using a *recorder device*, typically a laser recorder. The film is then developed, and the fiche production is complete.

Reading the data encoded on the Peviar film requires two steps. First, the fiche is captured by a camera or a scanner at a suitable resolution. The resulting image is then processed and the 2D barcodes are located and decoded by the *Peviar decoder*. The result of the decoding process is the original bitstream.

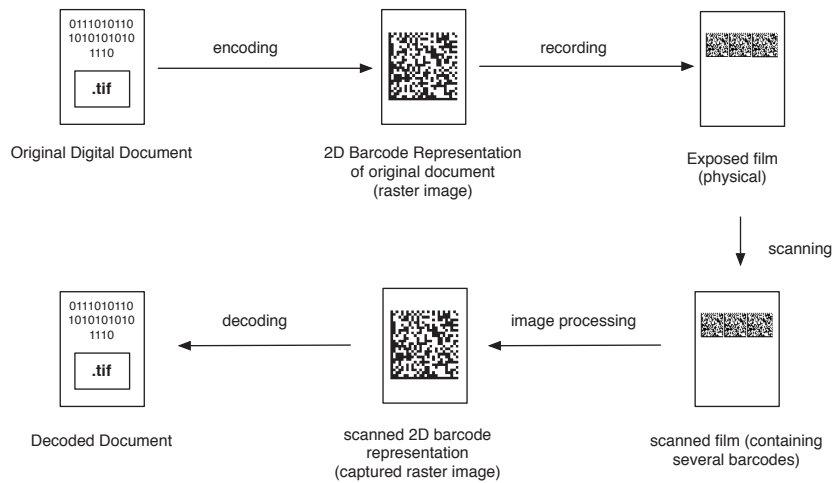


Figure 18: Illustration of the Peviar workflow (Image: F. Müller)

### 2.4.1 Peviar Channel Model

Writing to and reading from microfilm involves three imaging systems: an exposure device (the recorder), the film itself and a scanner (camera and optics). Figure 19 depicts an extended channel model of the process. It is evident that not only the film itself, but also the recorder and the capture device have an influence on the achievable information capacity.

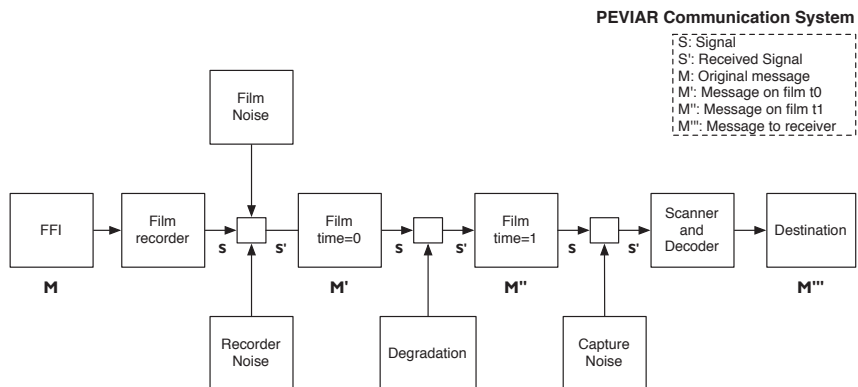


Figure 19: Storing and retrieving information on microfilm as a general communication system. It can be considered a three-channel process. First, the Film frame image (FFI) is recorded on the film. This recording is subject to the noise of the recorder and the film. Second, the film ages for a certain time, and is subject to degradation. Finally, the film is captured as an image, which introduces capture noise (Image: F. Müller)

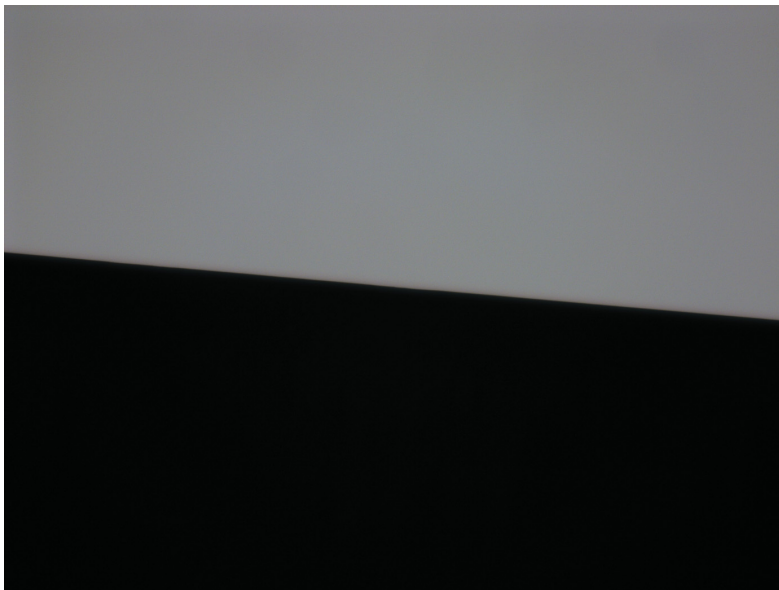


## 2.4.2 Modulation Transfer (SFR)

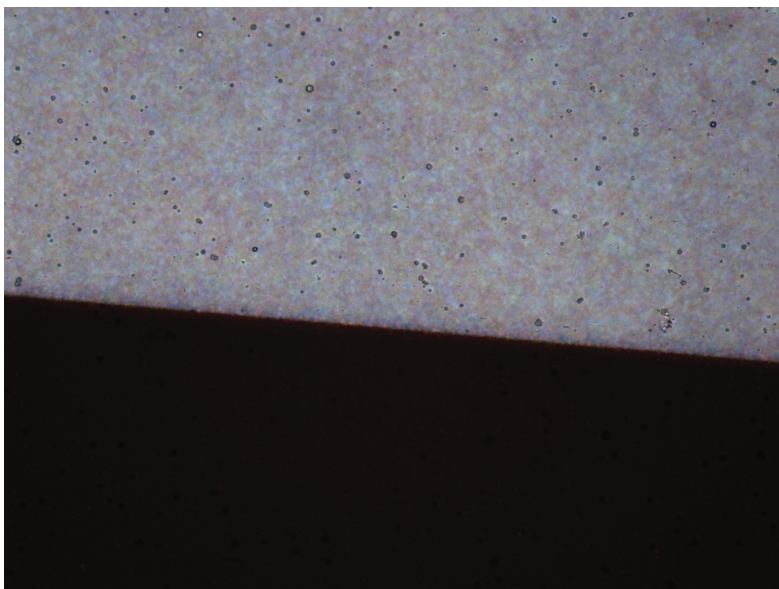
The spatial frequency response (SFR) measures the spatial resolving power of an imaging system. A measurement method for electronic capture devices has been developed and standardized under ISO 12233 [76]. While the procedure has been developed for electronic cameras, it is generally applicable to imaging devices. In the Peviar project, three imaging devices are involved: an exposure device (the recorder), the film itself and a scanner (camera and optics). It is important to note that each device has its own resolving power limit. In the following, the measurement method is briefly described. For a more detailed account, see Appendix A.1.

First, the target – a black to white edge – was placed under a Leitz Dia-plan microscope (numerical aperture 0.6). A camera mounted on top of the microscope (Zeiss AxioCam HR) was used to capture the target. A total of 8 captures of the target are required to perform one measurement. The target must be placed slightly slanted, at an angle of around 5 degrees to an orthogonal axis of the capture image plane. Four of the eight measurements capture the target with the white side left of the edge, four with the black side left of the edge. As a result of the capture process, eight images result. These images are then processed using the *sfrmat2* utility [77]. In order to reliably determine the SFR of a system, a total of eight SFR measurements must be performed (i.e. eight sets of eight captured images are processed). The eight measurements are aligned and their average is the resulting ISO-measured SFR of the system.

The measurement is based on a target exposed to film and read back by a scanner, the total system performance is measured – and not an individual component. However, in such a cascade, the total SFR results from the multiplication of the SFR of the individual components. In order to obtain the individual SFR measurements, the following was applied. First, a razor blade (see Figure 20) was placed under the microscope. The razor blade can be assumed to be infinitely sharp. Therefore, the SFR measurement of the razor blade captures measures the SFR of the capture device (microscope and camera) exclusively, we have *SFR Camera*. In a second step, an edge vacuum-deposited to the film (see Figure 21) was measured. This measures the SFR of the total of capture device and film (vacuum-deposition is assumed to be infinitely sharp, so it does not result in image degradation). Since we know *SFR Camera* from the razor blade measurement, we can simply divide the combined SFR through *SFR Camera* and obtain *SFR Film*, the SFR of the film material alone. Finally, an edge exposed to film (see Figure 22) is measured. This is the SFR of the total system recorder, film, scanner. The recorder SFR is determined by dividing the total SFR through *SFR Camera* and through *SFR Film*. Note that of course, this is only valid when using the same film and the same camera.



*Figure 20:* A razor blade captured by the camera. The blade is assumed to be infinitely sharp. The capture process introduces artifacts, since the high frequencies of the edge leads to anti-aliasing (Image: F. Müller)



*Figure 21:* A vacuum-deposited edge on film captured by the camera. There are no artifacts, since the camera is able to capture sharpness above the sharpness of the film (Image: F. Müller)

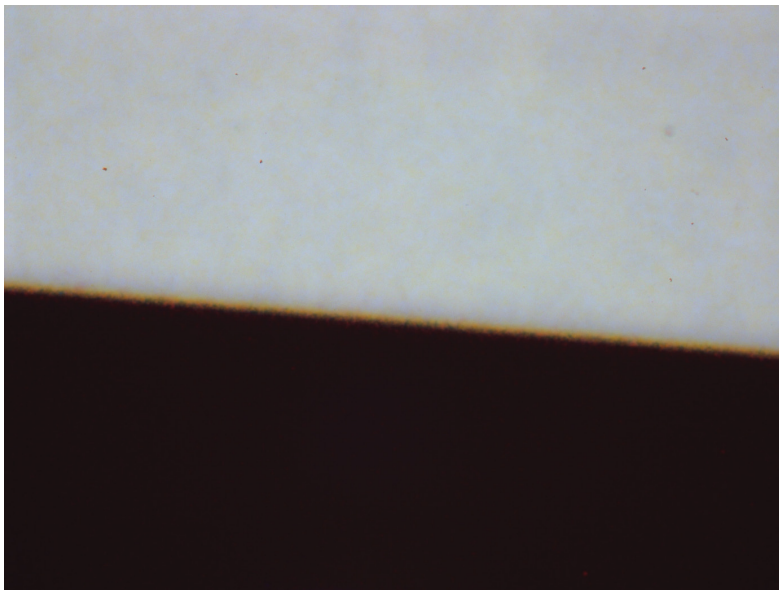


Figure 22: A laser-recorded edge on film captured by the camera (Image: F. Müller)

We are now able to determine the SFR of the total system and of the individual components. In the discussion of the information capacity of photographic materials, the relevant measurement was the modulation transfer function (MTF). The MTF is the traditional measure of resolving power of photographic material. SFR is the measure for electronic imaging systems. The two measures are sometimes used interchangeably and said to be the same<sup>34</sup>. The measures can be considered equivalent, with one notable difference: MTF is specified in line pairs per millimeter. SFR is specified in cycles per pixel, since the measurement is performed on a digital image, which in principle could have any real physical scale. The resulting cycles per millimeter are calculated by multiplying the measure by a factor which expresses the relation of an image pixel to its real physical size. In the case of our setup, one pixel has the size of  $\frac{1}{2410}$  millimeter, so the cycles per pixel measure was multiplied by 2410.

<sup>34</sup>Imatest, <http://www.normankoren.com/Tutorials/MTF.html>, accessed October 2011

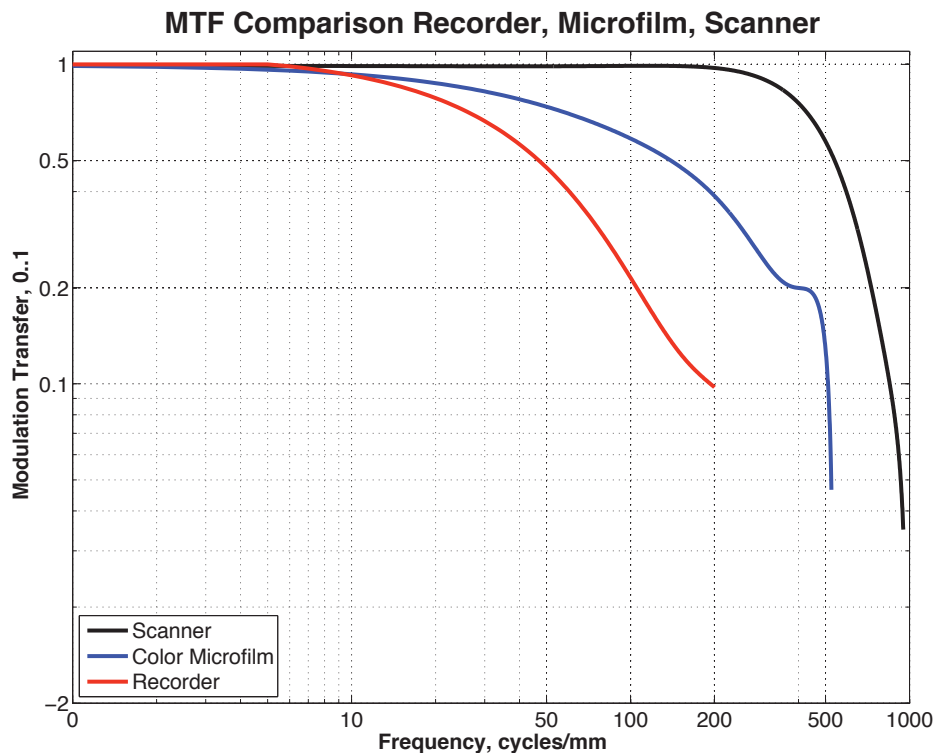


Figure 23: MTF curves for recorder (left, Fraunhofer IPM ArchiveLaser), microfilm (middle, Ilford CMM) and scanner (right, Zeiss AxioCam HR mounted on Leitz Diaplan microscope (NA 0.6)) (Image: F. Müller)

The MTF for all three components is shown in Figure 23. Clearly, the recorder is the limiting component. What we see in the measurement is the behavior of the modulation transfer with increasing spatial frequencies. The question, of course, is: which level of modulation transfer is relevant in determining the minimum raster point size necessary for reliable representation? In a first theoretical approach, we base our modulation transfer limit on the value of  $3dB$  (modulation transfer of approximately 0.7), which is the signal to noise ratio proposed by Shannon that is necessary to reliably transmit information through any given channel<sup>35</sup>.

<sup>35</sup>Using a relatively high limit, we consider ourselves to be on the safe side. In photographic applications, for example, the resolution limit is often set at a modulation transfer of only 0.1 or 0.2.

Cycles/mm	SFR
0.00	1.00
7.23	0.98
14.46	0.94
21.69	0.90
28.92	0.86
36.15	0.82
43.38	0.78
50.61	0.74
57.84	0.70
65.07	0.67
72.30	0.63
79.53	0.59
86.76	0.56
93.99	0.52
101.22	0.49
108.45	0.46
115.68	0.43
122.91	0.40
130.14	0.38
137.37	0.35
144.60	0.32
149.42	0.30

Figure 24: Tabular spatial frequency response report for microfilm as suggested in ISO 12233 showing the modulation response for various spatial frequencies. A spatial frequency of 60 cycles/mm results in a dot diameter of approximately 8µm (Measurement: F. Müller)

The table in Figure 24 shows the modulation transfer values for the recorder. We see that the response drops below the 0.7 (3dB) limit at around 60 cycles per millimeter. At such a spatial frequency, we could create dots with a diameter of approx. 8µm. Several exposure devices were tested<sup>36</sup>. They supported point sizes between 9µm<sup>37</sup> and 15µm. On a 600 meter 35mm color film roll, 25 gigabyte of data could be stored at 15µm point size<sup>38</sup>. At a point size of 12µm, 38 gigabyte could be stored. At 9µm, approximately 70 gigabyte would

<sup>36</sup>Fraunhofer IPM Archive Laser and Fluck Eternity 105. The MTF measurements published above are for the Archive Laser.

<sup>37</sup>The measurements show that the Archive Laser allows an 8µm resolution. However, its hardware and software only allow a dot size variance in 3µm increments.

<sup>38</sup>One frame 35 × 45mm has ((35mm\*45mm)/(15µm\*15µm)) / 8 = 875.000 bytes. Color microfilm has three layers, so a 600 meter roll has (600/0.045)\*3\*0.875 = 35.000 megabytes. We apply a fill factor of 0.9 (not all surface area can be used for data) and an encoding redundancy loss of 0.3, which gives us approx. 22 gigabyte. The other calculations are done in the same way.

fit onto the film roll. This amount of storage seems unable to compete with the amount provided by state-of-the-art storage technologies such as hard disks. But it should be kept in mind that in digital archiving, high data density is not always a critical factor, especially when measured up against stability and longevity.

### 2.4.3 Granularity (Noise)

The MTF measurements were done using a high contrast edge. This implies the use of two signaling levels and a binary code. But theoretically, the data density could be increased by representing more than two states with a single raster point. If we have 4 points at two possible gray levels (black and white), we can represent 4 bit. When using 4 possible levels, we only need two points to represent the same 4 bit. However, when using multiple levels of gray on the film, the granularity of the film gains relevance.

The granularity of the film gives a measure of how much variance in film density there is on a given, uniformly exposed area. Our measurement of the Ilford CMM shows that at very high and very low densities – basically, when working with black and white – its granularity is relatively small. But at medium densities - different shades of gray, which would encode a higher-level alphabet - the granularity increases considerably. These results could be expected given the nature of the Ilford film. At minimal optical density ( $OD = 0$ ), the dye is completely washed out, hence the granularity is zero. At maximum optical density, no dye at all is washed out, and the full dye remains in the film. Again, this results in zero granularity. In the case of partial dye wash-out, granularity does occur, and it reaches its maximum at medium optical density. Figure 25 shows our results. It can be seen that the granularity depends on the optical density: it is high at medium optical densities, and low at low and high optical densities. Figure 26 shows why this was to be expected. The observed property of the granularity means that when we use multiple levels of gray, we not only decrease the density distance between the levels - we also increase the medium-inherent fluctuation of a part of our dots. Let us interpret the granularity of the film as the noise component of a communication channel and the difference in density between the various levels as the signal strength. Then, we can say that by introducing further signaling levels, we not only decrease the signal strength, but we also and inevitably increase the level of noise on the channel.

In the RMS granularity measurement, several sections of the film surface that have been equally exposed are measured. The RMS granularity is the root mean square of the variance of the density for measurements in different areas. For a detailed description of the measurement, see appendix A.2. The size of the areas measured - the aperture - is an important factor. The variation is smaller for larger areas (see Figure 85 for varying apertures and their effect). The relation between the RMS granularity  $\sigma(D)$  and the aperture  $a$  has been described by Selwyn [78]. Selwyn showed that this relation applies to black and

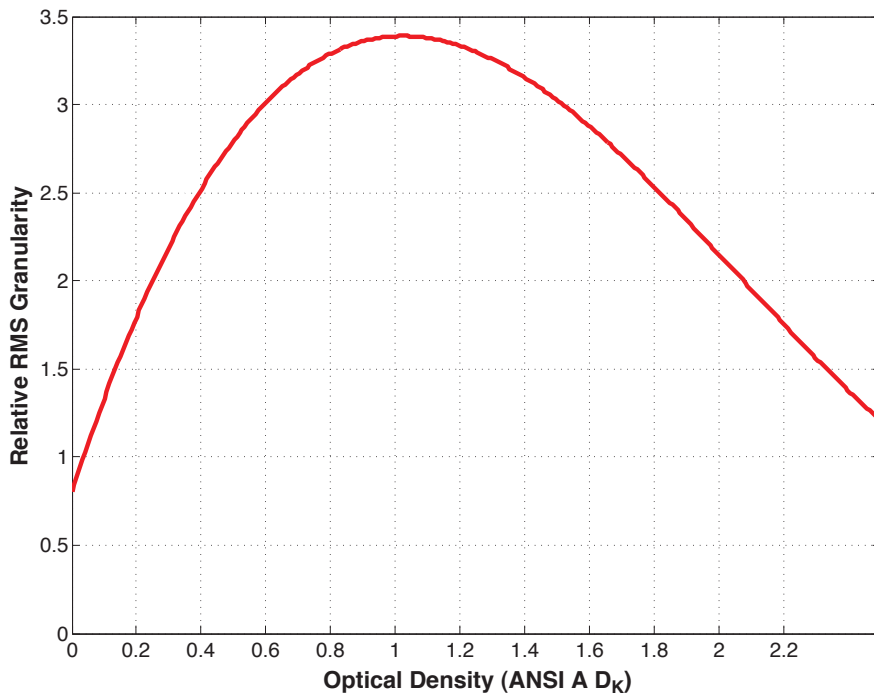


Figure 25: Relative RMS granularity for color microfilm (Ilford CMM). As expected, the granularity is considerably different from other materials, such as black and white film. It is evident that working with high and low densities results in dramatically lower granularity (noise) (Image: F. Müller)

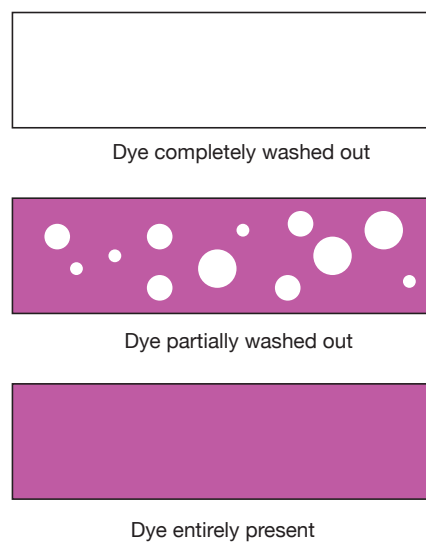
white film. We assume that it holds for color microfilm.):

$$\sigma(D) * \sqrt{a} = constant \quad (5)$$

When going from only two to more density levels, the signal strength is reduced (if we go from two to four density levels, the density difference between the levels is cut in half). This must be compensated by a respective reduction of the noise - if we only have half the signal strength, we can only support half the noise. Equation (5) shows that for a reduction of the granularity by factor  $n$ , an enlargement of the aperture by factor  $n^2$  is required. While the Selwyn relation is not applicable to microfilm across various density levels, but only for a given density level, it does suggest that the amount of data per area necessarily decreases as we use more density levels for representation.

We do not consider this to be a proof of the impossibility of increasing data density through the use of a higher alphabet. In the meantime, we have conducted experiments that support our intuition. In order to be able to properly detect the dot pattern when using more than two density levels, we had to increase the dot size. The loss of information density due to the increased area

was, in all scenarios, greater than the gain of information density due to the higher alphabet<sup>39</sup>. The effectiveness of a binary recording mode is also suggested by earlier literature mentioned before [74] [55]. Note that our discussion has focused on the optimum packing density of the film material alone. If we also consider the scanner and especially the recorder, we may come to different conclusions. If a recorder does not achieve the maximum resolution of the film (i.e. it has an inferior resolving power), a higher alphabet may be suitable. Determining the alphabet order can be supported by an experimental setup similar to the one used by Amir [79].



*Figure 26:* Schematic depiction of the magenta dye layer for different exposures. Top: high exposure leads to elimination of all the dye in the development process. Middle: medium exposure leads to partial elimination of the dye in the development process. Bottom: no exposure, no dye is eliminated in the development process. The reason for granularity in general is the varying size of the silver halides. In the case of the silver dye bleach process, this variance is only relevant when at medium exposures, i.e. when using the film (or a film layer) with shades of gray (Image: R. Gschwind)

#### 2.4.4 Error-Correction Codes

The key channel characteristics given by the bandwidth (SFR) and noise (granularity) determine the maximum achievable data rate. However, this maximum is only valid if the channel operates perfectly error-free. This, in practice, is not the case, for any real channel will be affected by errors. The errors of an error-affected channel can be compensated by the introduction of *forward error correction*. In forward error correction, the data that is to be sent over the channel

<sup>39</sup>We have concluded the experiment for 4, 8 and 16 density levels.



is encoded (sometimes called channel-encoded) using an error-correction code (ECC). The code adds redundancy to the data such that loss of parts of the (redundant) data can be compensated.

The simple example of *repetition code* illustrates this. Consider that a channel sometimes “flips” a bit, i.e. a '0' is sent, but a '1' is received, or vice versa. The frequency with which this occurs is called the bit error rate (BER). This can be compensated by repeating the message several times. In essence, the transmission is made redundant. Consider the example of a (5,1) repetition code: a single zero is sent as a sequence of five zeros, and a single one is sent as a sequence of five ones. On the receiver side, groups of five bits are *interpreted* as either a zero or one based on a majority vote: if more zeros have been received, the message is interpreted as a zero, if more ones are received, the message is interpreted as a one. The effect of such an encoding scheme is a dramatic reduction of the channel error rate. Table 3 illustrates this. The *Input BER* is the error rate of a single bit transmitted over the channel, the *Output BER* is the error rate of a single decoded bit (i.e. the bit obtained from sampling 5 single bits sent over the channel). Note that the added redundancy for a (5,1) repetition code is 400%. While this seems prohibitively inefficient, the amount of added redundancy depends on the channel error rate. A very reliable channel will have a low amount of redundancy, while a very unreliable channel will require a high amount of redundancy in order to be used reliably.

Input BER	Output BER
$10^{-2}$	$9.9 \cdot 10^{-6}$
$10^{-3}$	$1.0 \cdot 10^{-8}$
$10^{-4}$	$1.0 \cdot 10^{-11}$
$10^{-5}$	$1.0 \cdot 10^{-14}$
$10^{-6}$	$1.0 \cdot 10^{-17}$

Table 3: Post-decoding probability of bit errors when using a (5,1) repetition code (Table reproduced from [80])

As has been stated, every real channel is susceptible to errors. If a channel is to be robust, i.e. if we want to transmit information reliably, forward error correction is required. This is also the case for Peviar. Typical errors occurring on film material are scratches and dust. In the first implementation that was created, the well-established Reed-Solomon (RS) code [81] was used for error correction<sup>40</sup>. RS is a powerful code that is well-established, and for which encoding and decoding algorithms are widely available. Within the Peviar project, customized error coding has been investigated in-depth. The aim was to find an error coding scheme that was optimized for the special case of color microfilm,

<sup>40</sup>An open implementation is available at <http://www.eccpage.com>, accessed October 2011

thereby being more robust, efficient and powerful than non-specialized ECC. This work has been conducted by Ariel Amir at the University of Zurich. It should be noted that as of now, the specialized ECC has not been implemented in the Peviar workflow.

#### 2.4.5 Peviar Workflow and Specification

This section describes the Peviar workflow, i.e. the detailed steps required to store information on film, and to retrieve it. The process of storing information on film is called the Peviar write process, short writing, the process of retrieving information from film is called the Peviar read process, short reading. The processes are illustrated in Figures 29 (writing) and 30 (reading). Note that the processes are described from a technical perspective. Work items that are of crucial importance in the context of archiving, such as archival assessment, metadata generation and management, cataloguing, and others, are not considered. The goal of this section is to provide the reader with a complete description of the steps required to write and read information on Peviar microfilm. In both processes, the case of a single digital document ( $D$ ) being written or read is treated.

In principal, Peviar can be implemented on any visual medium, and in any format. The workflow is described with the example of the implementation that was used for the *Loveletters to the Future* project (see Section 2.4 and [58]). The base format is the 148 by 105 mm standard microfiche, which was used in portrait orientation. The fiche contains a header, where textual and image information is applied (around 20 mm). The rest of the fiche is reserved for the so-called *grid*, which splits the fiche into 6 (horizontal) by 7 (vertical) cells. A cell is either a metadata cell or a barcode cell. Metadata cells are used to place human-readable information such as images and text. Barcode cells are used to place machine-readable information in the form of two-dimensional barcodes.

**Peviar Write Process.** First, the bitstream of the digital document ( $BS_{org}$ ) is read and processed using the Reed-Solomon error correction code. The resulting encoded bitstream ( $BS_{enc}$ ) is one third longer than  $BS_{org}$ . It is the actual data that will be written, and its length is denoted by  $n$ . Each barcode cell has the same maximum capacity of data, and the total data must be split over as many barcode cells as necessary. The number of barcode cells required is calculated in the second step, and for each barcode block, a barcode with the respective part of the data is generated. No standard barcode is used, rather, a simple barcode algorithm developed at the Imaging and Media Lab is used. Common barcodes are optimized for robust applications: they have a high level of redundancy (possible damaging of barcode structure) and are often captured under non-optimal conditions (bad lighting, low-resolution cameras). Also, their capacity is very limited (the QR code, for example, has a maximum matrix size of 177 by 177, and a capacity of only several kilobytes [54]). The Peviar appli-

cation context differs in all these regards. The capture conditions are assumed to be good (good lighting, high-resolution capture device). Also, the barcode matrix in Peviar is much larger (several hundred to thousand cells per dimension, resulting in a capacity of tens or even hundreds of kilobytes). Therefore, using common barcodes for Peviar would be inefficient. The custom Peviar barcode has two important features: parity checking and error delocalization. Groups of eight bits (a byte) are taken, and a parity bit is calculated (if the number of zeros is even, the parity bit is zero, if the number of zeros is odd, the parity bit is one). These nine bits are stored in a three-by-three table with the parity bit in the center and the bit count starting top left and going clockwise. The encoding of an original bitstream is illustrated in Figure 27. Such three-by-three bit groups are called *parity blocks*. The parity blocks are not placed one next to the other starting from top left, but are distributed within the barcode plane using a pseudo-random algorithm. The reason for this is that typically, damages to the film material (scratches and dust) are spatially concentrated. For any given sequence of bits, only a certain error rate can be compensated by ECC. If the spatially concentrated damage affect a consecutive sequence of bits, the error rate becomes too high locally, and a part of the bitstream is lost. Therefore, the parity blocks are randomly spread over the barcode plane, placing consecutive bytes spatially separated. This is called error delocalization: the spatial and byte-order location of the error are made distinct. The barcodes are provided with two control marks at every corner in order to enable detection of the barcode block during reading. Also, a so-called *alignment border structure* is put alongside the edges of the barcode. This will allow the identification of the barcode lines and columns during reading (see Figure 28).

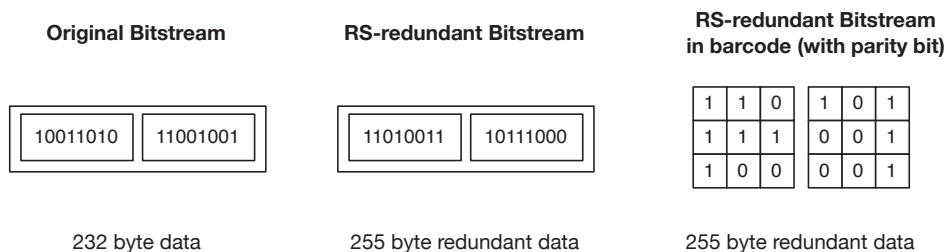


Figure 27: For every 232 bytes of the original bitstream, a 255 byte block is created. The added 23 bytes hold the redundancy. Every byte of the error-encoded bitstream are placed in a barcode module. Each module contains the 8 bits of the byte and a parity bit for error detection.

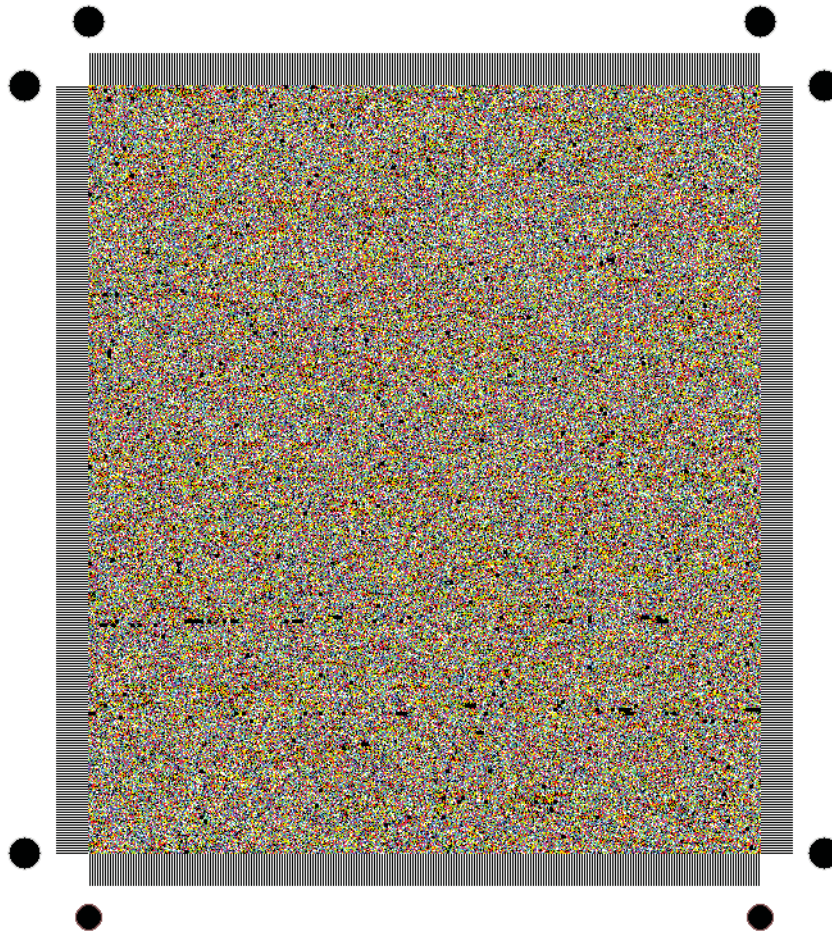
Once all the barcodes have been generated, they are saved as raster images. Depending on the spatial resolution chosen, one logical barcode entry corresponds to one pixel, or to a x-by-x pixel square. This measure is called logical pixel size (LPS) and is expressed in physical pixels. In the LLTF project, an

LPS of 6 was used, so every barcode bit occupies 6 by 6 pixels, and a parity block occupies 18 by 18 pixels. In the third step, which can happen prior to, in parallel to or after step two, the contents of the metadata cells are prepared in the form of raster images. Once all the barcode and metadata cell images are prepared, the fourth step assembles the image(s) that will be exposed to the microfilm. This image – or these images, if not all cells fit onto one fiche – is called film frame image (FFI)<sup>41</sup>. In the fifth and final step, it can be exposed to the film, and the Peviar write process is complete<sup>42</sup>.

---

<sup>41</sup>In early project stages and in the DANOK project, rolls of 35mm film were used to record information. These rolls have frames as their subdivisions, which is why we speak of film frame images, rather than film fiche images

<sup>42</sup>Of course, any responsible archivist would not consider data to be written unless it would have been proof-read. But as stated in the beginning, this process description is minimal and should only explain the basic technical steps



*Figure 28:* Individual barcode cell. The border alignment structure helps in detecting the codes rows and columns. The barcode has three color layers (one for every layer of the microfilm), the red-, green- and blue-components of the image address the red-sensitive, green-sensitive and blue-sensitive film layer, respectively (Image: F. Müller)

**Peviar Read Process.** First, the fiche is scanned and made available as a raster image for software processing. As suggested by the Nyquist-Shannon sampling theorem (see concisely in [82], p. 837), an oversampling of at least factor 2 is required to sample a signal without generating aliasing. In practice, we have worked with an oversampling factor of 3. With this factor, the processing software has proven to work much more reliably than with the theoretical minimum oversampling. In the second step, all barcode cells are located. This is achieved by using the control marks applied to the corners of the barcode. In case the scanned barcodes are not orthogonal, they are made orthogonal using an affine transform. In the third step, the alignment border structure is analyzed in order to estimate the positions of the individual barcode pixels

(logical pixels) within the barcode plane. In the fourth step, the bit pattern of the original barcode is reconstructed using the estimated barcode logical pixel positions and sampling them in order to determine whether they are a zero or a one. In the fifth step, the barcode generation of the write process is actually reversed. Using the pseudo-random algorithm used for delocalization, the original sequence of the parity blocks is assembled. Every parity block is checked, i.e. the parity calculation is repeated based on the sampled bits. If the parity check is successful, the byte of the parity block is added to the parity block sequence. If the parity check fails, the failure is marked. The result of reading all the parity blocks in the right order is the estimate of  $BS_{encoded}$  of the write process  $BS'_{encoded}$ , which may contain marked parity failures and erasures (wrong bits not detected by parity check). In the sixth step,  $BS'_{encoded}$  is processed using a Reed-Solomon decoding algorithm. The algorithm will detect any erasures and attempt to repair them. If the reconstructed bitstream has an error rate smaller than redundancy, the algorithm will provide the original bitstream ( $BS_{orig}$ ) as a result. Otherwise, it will send a notification that the original bitstream cannot be reconstructed. In the seventh and final step, the decoded bitstream is stored as a file, and the originally stored document is available.

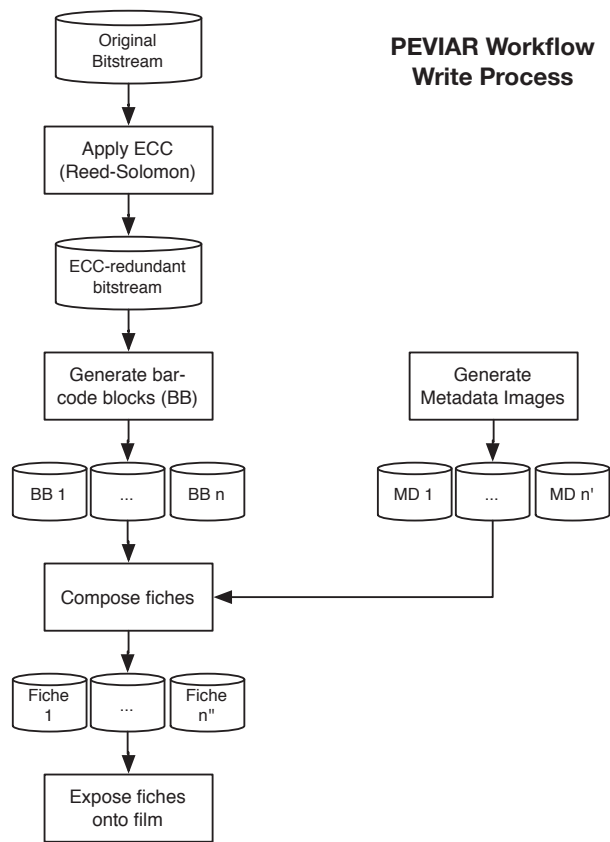


Figure 29: Peviar write process (Image: F. Müller)

**PEVIAR Workflow  
Read Process**

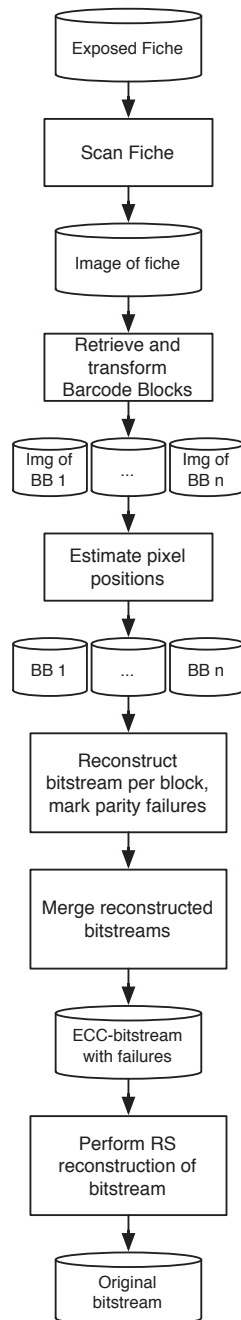


Figure 30: Peviar read process (Image: F. Müller)

In the previous sections, the technical foundations for storing bits on film, as well as the respective workflow have been presented. In this final section, we attempt to derive a specification of Peviar. It is not a direct result of the



maximum achievable data density, as could be expected. Rather, it is derived from considerations about the read process, more precisely, from the scanning part. In order to be able to read back a Peviar fiche, it must be captured at a resolution that allows reliable detection of the relevant features, which has been shown to be at least twice the actual resolution (the theoretical minimum), or better three times the actual resolution. If the physical pixels representing data are as small as possible (around  $8\mu m$ ), this means that the scanning hardware must provide a very high resolution in the area of between 2 and 3  $\mu m$  only. Such scanning hardware is not only expensive, it may also be unsuitable for mass processing.

The main idea of Peviar is the preservation of information. We have seen how this is solved on a technical level, however, the technical level is only one level of concern. A very important aspect in information preservation are costs – while the preservation of vast amounts of information is technically possible, it may still be omitted for reasons of prohibitive costs. Peviar is technology independent in that in terms of hardware, it only requires a high-resolution image capture device such as a camera or scanner. However, if a special capture device would have to be constructed in order to efficiently capture the fine details of a Peviar fiche, the main advantage would have been lost. Therefore, the feasibility of read-back and the costs associated with it are of crucial importance.

The following table proposes the classification of Peviar into three categories: laboratory, professional, and consumer. This is not a hint for a commercialization of the Peviar technology, but stems from the classes of scanning devices that exist. Laboratories have binoculars and microscopes to detect very fine details contained in nature's wonders and marvels. They are designed to provide the greatest detail possible, and not to cover great amounts of space. Image capture professionals have high-end cameras and scanners in order to provide supreme quality captures of all sorts of objects for customers. They do not reach the resolution of a microscope, but can cover a much greater area. Finally, image capture used in every-day life, such as by consumers, can be seen as a less pricey, less powerful version of professional image capture devices: they can cover the same area, but not with the same level of detail.

*Table 4:* Peviar technical specification for several resolution classes. A provides maximum data density, C provides minimum readback hardware requirements, B is a compromise.

	<b>A class</b>	<b>B class</b>	<b>C class</b>
<b>LPS <math>\mu m</math></b>	9	50	150
<b>Megabyte/microfiche</b>	15	0.5	0.05
<b>Fiches/1MB</b>	$\frac{1}{15}$	2	20
<b>Scanning DPI</b>	85.000	15.000	5.000

## 2.5 Authenticity and Originality in the Digital Archive

In any archive (digital or not), assessment of document integrity and authenticity is important. Questions in this regard have been relevant in the archiving community for over 100 years [83]. We have seen before that regarding information preservation, the digital archive is fundamentally different from the analogue archive. Regarding authenticity, the difference is also considerable, albeit some important similarities remain. This section first discusses authenticity and integrity in analogue and digital archives. It will be shown that although there exist methods for ensuring authenticity in the digital domain, some problems remain unsolved. The unique properties of Peviar, which can be considered a paper-equivalent for the digital world, make it interesting as an answer to questions of authenticity in the digital archive. First, however, we will clarify some concepts.

According to recent literature in the field, a document<sup>43</sup> possesses integrity if compared to its original state, it has not been altered at all, or, if any alteration is documented and known to the consumer of the document. In terms of digital documents, this means that the bitstream of a document remains exactly the same. The authenticity of a document is established if the document actually is what it purports to be (see [84], [85], [86] for different perspectives on integrity and authenticity). The ISO Records Management Standard [87] states that an authentic record is “one that can be proven (a) to be what it purports to be, (b) to have been created or sent by the person purported to have created or sent it, (c) to have been created or sent at the time purported”. Integrity and authenticity are aspects of documents sometimes discussed separately. In this work, integrity is regarded as a partial feature of authenticity. This is justified if we consider the first condition (a) of the ISO definition of an authentic record. The link between what a document purports to be and what a document actually is demands that the document possesses integrity – in case integrity is lost, so is authenticity, since the document no longer *is* what it purports to be.

As evident from the definition, there are various factors that have an impact on document authenticity. Suppose that a manuscript is discovered in the literary remains of a writer. What would it mean for such a manuscript to be authentic? Obviously, we cannot answer this question so far: for we do not know what the manuscript purports to be. Now suppose that someone claims that this is the long-awaited, never-found novel *La Última* about whose existence there had been much speculation. Now we can verify whether the manuscript is authentic, for it purports<sup>44</sup> to be a document created by the author in question, at a time during his life, intended as fictional work (the intention is interpreted as the *what it purports to be* part). We can imagine

---

<sup>43</sup>Authenticity is sometimes discussed regarding electronic records. In this work, we only speak of digital documents, which can be considered a more general class than electronic records (this would be clearer if we would use the term *electronic documents*)

<sup>44</sup>For a perspective on who or what actually makes the claims, see [83], p. 6

that in such a scenario, the entity wanting to verify the authenticity of the document will do so by consulting various experts, such as technicians that are able to carbon-date paper and ink, calligraphists for comparing the handwriting with other (*authenticated*) handwritings of the author, and a man of letters for giving an assessment of the literary content. If the expertise permits it, the document can be successfully authenticated. Suppose further that there is a foundation devoted to the management of the writer's inheritance. It has an archive where it would like to keep the manuscript. After it has been authenticated, the manuscript is introduced into the archive. If the manuscript is kept without care, someone may be able to manipulate it (e.g. remove a page, or add text). Such manipulations – changes to a document that are not made explicit and that result in a loss of information – affect the integrity of a document and thereby also its authenticity. This means that once the authenticity of a document has been assessed, it cannot be taken for granted – it must be actively safeguarded. The foundation archive, aware of this, will have procedures in place to permanently assure its documents *remain* authentic. Suppose that it does so by ensuring that access to the manuscript is strictly controlled and limited, and that a manipulation of the manuscript is not possible (the archivist is a trustworthy person, and the manuscript can only be seen under his supervision). The chances are good that given these measures, the manuscript will remain an authentic document in the assets of the archive.

### **2.5.1 Problems of Authenticity in the Digital Archive**

In order to understand the problems with authenticity in the digital archive, the example of the previous section is modified. Suppose that the writer has used a computer to write his books, and that instead of a manuscript, a floppy disk is found in the writer's literary remains. This floppy disk contains a text file (furthermore called digital manuscript) that someone claims is the long-expected, but never delivered novel *La Última*. How can this be verified, i.e. how can the digital manuscript be authenticated? Actually, it is not so different from the analogue example, even though different experts may be involved. A computer expert could try to validate if this floppy disk was actually used by the writer (she could look for traces of it on the computer, looking for mentions of it in electronic communication etc.). No calligraph would be required, however, a man of letters could still assess the literary content of the text. Suppose that the experts agree that the digital manuscript is in fact the last novel of the writer. After authentication, it could be accepted by the foundation archive. In the analogue archive, a sound way to safeguard the future authenticity of a document is to restrict access to it and leave it untouched. If this is done in a digital archive, the consequences might be catastrophic: material decay of data carriers and hardware and software obsolescence require an active preservation strategy, and this strategy usually involves operating on the documents in one way or another. Making sure that documents remain unaltered by limiting and

strictly controlling access to them is a preservation strategy that is unlikely to be successful in the digital archive. In a sense, document alteration becomes a necessity in the digital archive, and this alteration is a great threat to authenticity. Suppose that the manuscript must be transferred to a new storage medium five years after having been archived. A new document is created<sup>45</sup>, its authenticity is derived from the relation and comparison to the originally archived manuscript. The problem, of course, is not that we now have a different document, but that in any migration process mistakes can be made. Copying and converting data can be erroneous (conceptually and technically) and allows the introduction of willful and accidental manipulations. Furthermore, the integrity of digital documents is threatened at another level. Most media that store digital data are rewritable (e.g. magnetic media). A document present on a carrier can be altered in any form, and any alteration is unnoticeable (it may be through in-depth physical analysis of the carrier, but this would require disproportional resources for most documents). Clearly, this is different for our analogue manuscript, where any alteration will leave physical evidence that may be detectable through various forms of analysis (forensic, calligraphic etc.).

In the scenario just described, there are actually three problematic aspects. At the time of authentication, we lack material evidence such as handwriting that make documents unique and distinguishable from documents from other sources. During preservation, we have the necessity of migration, which implies that we must copy the document every once in a while. Also, on some storage technologies, manipulations are possible and very hard to trace by default. All these aspects have one common root: the lack of materiality of digital documents – *digital documents have no relevant materiality*.

### 2.5.2 Cryptographic Techniques

Several techniques exist to ensure authenticity and integrity in the digital archive. Most of them are based on procedures originating from cryptography [88] [89]. In the following, two techniques are described: hash summing and digital signatures.

In hash summing, a *hash sum* (also called checksum) is computed for a document. A hash sum is a datum computed from the data of the digital document. The function responsible for computing the hash sum is the hash function. A hash function  $h$  has the following properties [82]:

1. For any amount of data  $M$ ,  $h$  is able to compute a hash value  $h(M)$  of fixed length
2. It is impossible to compute  $M$  from  $h(M)$

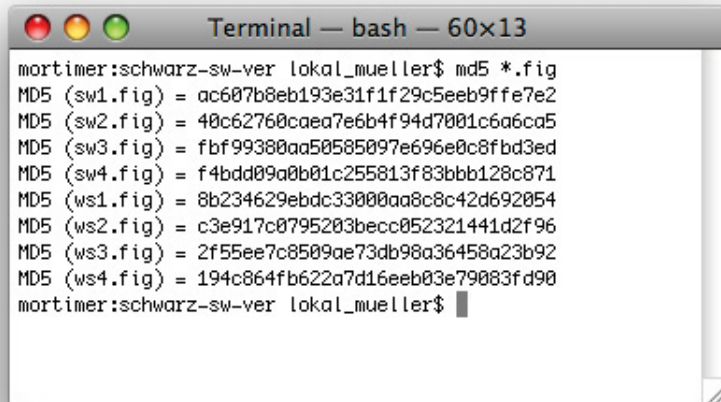
---

<sup>45</sup>In a sense, it is not a new document – the content should be the same. But physically, we have a new document – a new series of zeros and ones on a different medium.

3. It is practically impossible to find two sets of data  $M$  and  $M'$  such that  $h(M) = h(M')$

The impossibility in (3) is not absolute, since the infinite space of all possible data is mapped to the finite space of values of a fixed length. The case of  $h(M) = h(M')$  is called a collision, and the probability of the collision depends on the length of the hash value. The long-established MD5 (Message-Digest 5) has a hash length of 128 bits. Today, this is considered unsafe, as it has been shown that creating a collision is feasible [90]. By using longer hash values (e.g. 256 or even 1024 bits), the probability of collisions decreases, and the reliability of the hash sum increases.

Given that they are sufficiently reliable, hash sums can be used to identify documents regarding their bitstream. Changing one single bit of a document is required and sufficient for changing its hash sum. Hashing is a simple, yet effective means of tracking document integrity. Using tools such as the *md5* command line program (see Figure 31), hash sums can be calculated for any files desired. The hash sums are typically stored in a file alongside the original documents. At any given time, the integrity of these documents can be verified by re-computing their hash sums. If they match the original hash sums, the documents possess integrity.



```
Terminal — bash — 60x13
mortimer:schwarz-sw-ver lokal_mueller$ md5 *.fig
MD5 (sw1.fig) = ac607b8eb193e31f1f29c5eeb9ffe7e2
MD5 (sw2.fig) = 40c62760caea7e6b4f94d7001c6a6ca5
MD5 (sw3.fig) = fbf99380aa50585097e696e0c8fbd3ed
MD5 (sw4.fig) = f4bdd09a0b01c255813f83bbb128c871
MD5 (ws1.fig) = 8b234629ebdc33000aa8c8c42d692054
MD5 (ws2.fig) = c3e917c0795203becc052321441d2f96
MD5 (ws3.fig) = 2f55ee7c8509ae73db98a36458a23b92
MD5 (ws4.fig) = 194c864fb622a7d16eeb03e79083fd90
mortimer:schwarz-sw-ver lokal_mueller$
```

Figure 31: Use of *md5* tool to create hash sums for documents. The hash has a length of 128 bit and is output as a 32-digit hexadecimal number (Image: F. Müller)

The problem with hashing is that the original hash sums – the ones that will be used to verify document integrity – are typically stored as digital documents themselves. In principle, they can be arbitrarily manipulated, and a willful manipulation of a digital document could go alongside with the replacement of the

original hash sum with the hash sum of the manipulated document. A hash sum could be used to verify the integrity of the document containing the hash sums, but then, the problem is recursive and has no fundamental solution. The most secure way for storing the hash sums would be paper. But on paper, the hash sums would not be available as digital documents and could not (without preprocessing) be used for automated document verification. Also, hash sums only help in determining whether two documents are identical at the bitstream level. They say nothing about content [89], and the only information that a hash sum comparison can yield is that two documents are not identical – it has no information about the nature of the manipulations that have taken place.

Finally, signature techniques can be used to verify document authenticity. They are based on asymmetric cryptographic protocols. At the core of asymmetric cryptography, we have a pair of keys, one called *public key*, one called *private key*. These keys have the following relation:

1. A document encrypted with the public key can only be decrypted with the private key
2. A document encrypted with the private key can only be encrypted with the public key
3. The private key cannot be calculated from the public key (and v.v.)

The two keys of a public/private key pair complement each other. Anyone participating in communication secured by asymmetric protocols must first obtain such a key pair. These keys can be generated by one self, but are usually provided by a trusted certificate organization – more of that later. It is imperative that the private key is not disclosed to anyone – disclosure would make the protocol insecure. The public key, on the other hand, must be disclosed to the public. Asymmetric cryptography has two fundamental applications: secure transmission and signing. They are illustrated in Figure 32. In the case of secure transmission, the sender will obtain the public key of the recipient. She then encrypts the message using that key and sends the encrypted message to the recipient. The message can only be decrypted with the private key of the recipient. The protocol thus assures that the message can only be read by the recipient. In the case of signing, the sender of the message encrypts it using her own private key. The message can then only be decrypted using the public key of the sender. The recipient of the message then verifies whether it is from the claimed sender by attempting to decrypt it with the claimed sender's public key. The authenticity is then implicit: since the message can be decrypted using the public key of the sender, that sender must have encrypted it using her private key.

## Two Applications of Asymmetric Cryptography: Secure Signature and Secure Transmission

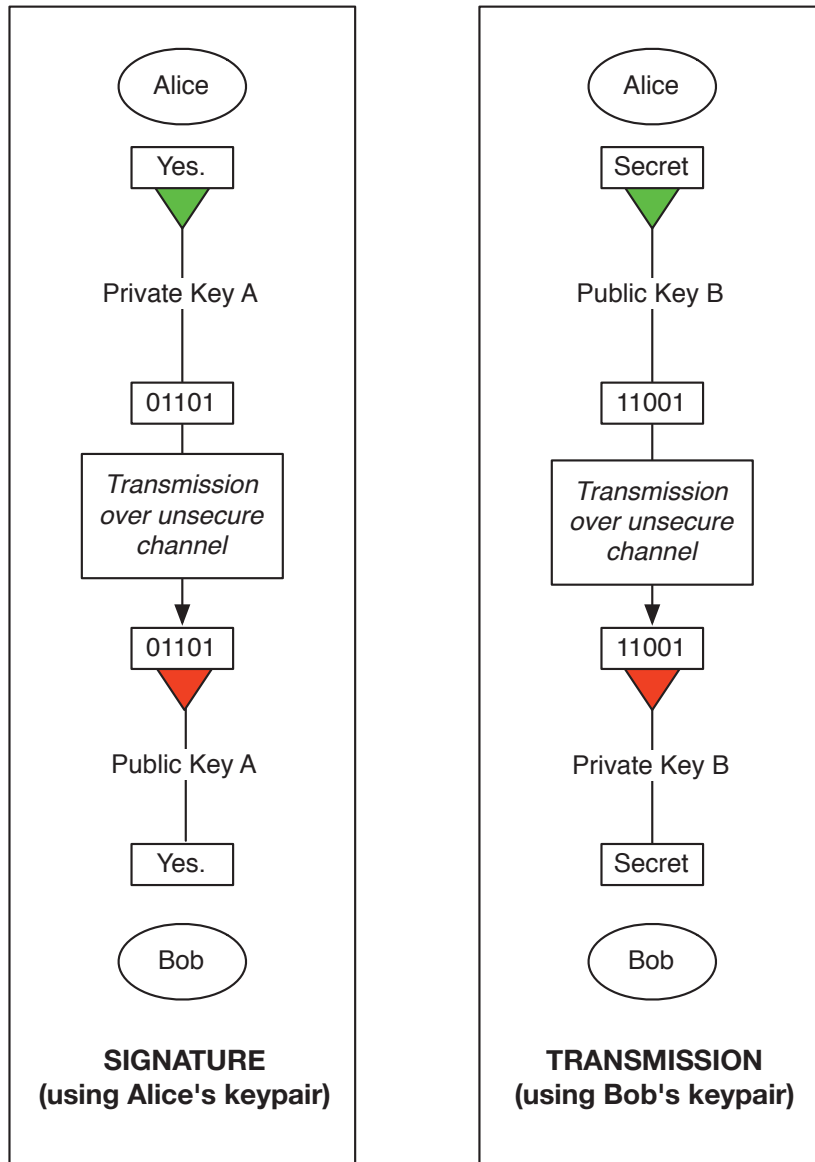


Figure 32: Applications of asymmetric cryptographic protocols for the partners Alice and Bob. Signature (left) enables a sender to sign her messages, decrypting it with her private key. The recipient can verify the message signature by attempting to decrypt it with the public key of the sender. Transmission (right) allows secure transmission without key exchange. The sender encrypts the message using the recipient's public key, once encrypted, the message can only be decrypted using the recipient's private key (Image. F. Müller)

In digital preservation, it is generally considered that preserved documents must not be encrypted ([89], p. 184). The main reason is a possible loss of the key: in such a case, the document would be lost<sup>46</sup>. Therefore, not the original document, but a related document is signed. Boudrez describes the two-step method of signing digital documents with an “advanced digital signature” ([89], pp. 180ff). In the following, a complete signing and authentication procedure is described for the Sender  $S$ , the Recipient  $R$ , the document  $M$ , and the senders key pair  $K(S)_{private}$ ,  $K(S)_{public}$ .

1.  $S$  computes the hash  $A$ ,  $H(A)$ , and stores it as document  $B$
2.  $S$  encrypts  $B$  using  $K(S)_{private}$
3.  $B$ , a digital object distinct from  $A$ , is now the digital signature of  $A$
4.  $A$  and  $B$  are sent to  $R$
5.  $R$  receives  $A'$  and  $B'$
6.  $R$  decrypts  $B'$  using  $K(S)_{public}$
7.  $R$  computes the hash of  $A'$ ,  $H(A')$
8. If  $H(A')$  equals the decrypted  $B'$ , the document is authenticated
9. Otherwise, either the transmitted document or the signature has been compromised ( $A \neq A'$  or  $B \neq B'$ )

This is probably one of the more advanced and error-proof methods of using cryptographic techniques to ensure authenticity. While it does offer the possibility to tell whether a document is authentic and unaltered, it would not be of much help in case a manipulation is determined: we would have no clues as to how the document was altered. There are more severe challenges, however. It was stated initially that certificates are usually provided by a trusted issuer. They have a limited validity (usually one year), and their validity depends on the issuer guaranteeing it. This is a problem for long-term preservation. How can we ensure that the issuer of the digital certificates used to authenticate our archives of today will still exist tomorrow? We cannot. It is important to note that the reliability of the asymmetric protocol does not only depend on its computational aspects, but also on the manner in which keys are created, issued and managed. The safest way, according to current standards, which uses authoritative issuers, is not an option for digital preservation.

---

<sup>46</sup>Cryptographic protocols have the aim to make decryption of encrypted information impossible without the proper keys. Although over time, many protocols have been broken (thus allowing information decryption without the key), this is not something one should rely on in digital preservation.



Cryptographic techniques are certainly a valuable tool in the authentication of digital documents. But very similar to the problem of preservation proper, they are not a solution that is ready for the future. First, increasing computational power gives the implementation of cryptographic protocols only a limited lifetime. We have seen the example of hash computation, where 128 bit has values have recently been proven to be too weak to be reliable. More generally, cryptography reacts to progressing compute power by increasing key sizes<sup>47</sup>. Also, the involvement of certificate organizations that must continue operations (and be paid, probably) in digital signatures introduces a dependency not resolved at the time of storage.

Besides cryptographic techniques, social techniques to ensure authenticity continue to exist in the digital domain. Take the case of documents archived on a re-writable medium. While this in principle allows arbitrary manipulation of data and metadata (e.g. of documents and their signatures or hashes), this can be countered by implementing a strict security protocol on the organizational level, e.g. four-eyes principle for both physical access and migration procedures. And they probably will employ protocols for ensuring the authenticity of their digital document. Trust in institutions – rather than only in material artifacts – is an aspect that will continue to be important in the time of digital archives: as a user of an archive, for most purposes, it will suffice to rely on the trustworthiness of the institution [91]. However, the underlying technical problem remains: for the institution actually concerned with the preservation of the digital document, ensuring its authenticity still means having to deal with the problems presented by migration, the principle possibility of manipulation and difficulties associated with any cryptographic technique.

### 2.5.3 Peviar: Digital Originals

It has been stated that the main problem of authenticity in digital archives comes from the fact that digital documents lack relevant materiality. They are not strongly and irreversibly bound to a material carrier, but are rather only temporarily captured by physical media. Throughout their life in the archive, digital documents are migrated to ever new carriers, and are susceptible to accidental or willful manipulation at times of migration, and possibly also in between. In this respect, Peviar offers a solution. It will be shown that authenticity is provided by the absence of migration, the impossibility of manipulation, and the difficulty of duplication.

Peviar is (virtually) migration-free. Digital documents archived with Peviar are not required to migrate, hence, the document of the future and the document of today are one and the same. In addition, the properties of developed photographic material are such that manipulation of documents stored with Peviar is impossible. Once film is exposed and developed, the image contained on

---

<sup>47</sup>DES (Data Encryption Standard) 1975, 56 bit key. AES (Advanced Encryption Standard) 2000, 128-256 bit key

it is made permanent. This process is irreversible. As soon as it is completed, the image on the film cannot be altered any more. Any manipulation would result in visible damage. Thus, any alteration attempt would be detectable, and through the use of forward error correction in the barcode, it could be completely compensated up to a certain level of damaging. Finally, Peviar microfilm is very hard to duplicate. The film is exposed from an image containing both the barcodes and analogue information (text and possibly images). In order to create a perfect copy of the film, the original exposed image must be available. Attempting to duplicate it with the exposed film as a master (scanning and exposing, or contact copy) will result in a loss of resolution. This would be noticeable in the analogue part of the information, and it would also severely affect the barcodes, possibly rendering them unreadable. Even if the original master image used for exposure is available, the exposure itself is not done easily. Only specialized laboratories are able to expose color microfilm at such high resolutions. In short: if digital documents are to be manipulated by introducing a manipulated duplicate of the original film, this is very hard to achieve, and it would probably still be detected.

The term *original* may not be applicable to digital documents per se. Instances of digital documents are, without any further ado, indistinguishable amongst one another. Even if we consider one instance of a document present on a specific medium, it usually does not make sense to speak of a possible original: the document can be altered, copied and moved at no great cost. In the case of Peviar, the bond between the digital document and the physical data carrier is much stronger. It cannot be altered or moved, only damaged or destroyed. It can be copied, but creating an (almost) identical copy means replicating the microfilm, which is hard for reasons mentioned. Peviar microfilm is as close to a *digital original* as one can get. It allows digital documents to be created as original artifacts that guarantee the prolonged availability and authenticity of the information they preserve.

**Part III**

# **Evaluation - Harvesting a Social Graph**



### 3.1 Social Computing: Theory and Current Practice

In 1982, the Time magazine made an unusual choice for the recipient of its *Person of the Year* award. Instead of an individual or a group of individuals, *The Computer*, i.e. the personal computer, was given the honor. It marked the introduction of a technological product that would have a significant impact. Only 24 years later, in 2006, the award was given to *You*, shorthand for the (anonymous) group of internet users contributing to the web with their user-generated content in the form of blogs, Wikipedia articles, forum entries, pictures and videos. The applications of what can most generally be termed *Web 2.0* have not only made the web read-write, allowing anyone to consume and produce information. It has also made it a place where social interaction takes place. Instead of the machine, the user has become the center of attention. Ever since the rise of the Web, the formation of online communities could be observed. While earlier tools such as messageboards and Wikipedia have resulted in the building of online communities centered around a specific topic or activity (we could say that the online community was *original*), more recent dedicated social networking tools such as MySpace, Facebook and LinkedIn enable users to project their existing (real) social network into the online world. Friend- and kinship is no longer either virtual or real. Rather, it can be expressed and lived in both real and virtual spaces. The *social media* technologies that enable this transition are the foundation of the concept of *social computing*.

Wang et al. [92] recapitulate several definitions of social computing. At the most general level, social computing can be described as “any type of computing application in which software serves as an intermediary or a focus for a social relation” ([92], p.79). Furthermore, social computing introduces the notion of the *social context*. The social context of an individual is the sum of all relationships that it has to other individuals, including information on the nature and quality of these relationships. This is illustrated by the services Facebook (and other social networking sites) provide to third-party websites. Visitors of, say, a newspaper website have the ability to authenticate their identity via the social network. This gives the newspaper the opportunity to link visitors, should they actually have a relationship in the social network. So now, when browsing local or global news, one knows which articles have been “liked” or “disliked” by friends and acquaintances and is, so to speak, always embedded in one’s social context. In a more systemic and less user-centric view, social computing focusses on interactive networks – the base unit of social computing is not an individual user or an individual device, but rather a conglomerate of users, devices and services that contain rich interlink information.

Social media applications have had a very successful development in the past years. Wikipedia, the collaborative online encyclopedia, has grown into one of the worlds largest collections of knowledge within ten years since its foundation in January 2001, providing over three and a half million articles in its English version alone [93]. Facebook, within seven years since its foundation in 2004,

has grown into a network of individuals and organizations with over 500 million active users spread over the globe [94]. On Twitter, a short-message service inspired by the mobile phone short message service (SMS) and following a publisher and subscriber (*follower*) model, 90 million messages were published daily in July 2010 [95], connecting users with topics and hyperlinks. Social computing applications have not only been very successful. More and more, they are being adopted universally. MySpace, an early social media site with a strong focus on music and music culture, was predominantly adopted by younger users. The same was true for Facebook, whose first target demographic were university students and recent graduates. Today, the largest demographic group on Facebook are users between 45 and 54 years of age, which make up 26% (data taken from Google AdPlanner [96], Google account required). In an overview of social media sites, a blog author shows that nearly every demographic group is well-represented (or even predominant) on one of the many different sites ([97], data is taken from AdPlanner). As a witness to the growth and expansion of social media, one could interpret it as just another successful internet technology, comparable to the impact of peer-to-peer services (which are also social, but focus more on content delivery rather than communication). But seen as a collection of tools that revolutionize the way people communicate, social media are part of the more general field of social computing, whose impact on our society is yet to be determined.

In a recent editorial, Riedl [98] suggests that the tools that we use to communicate and the possibilities they offer have a strong impact on our communication – given new tools, we will not continue to do the same things we did before, but the new possibilities will actually fundamentally change the way we do things. He follows up on a thought developed by Dunbar, who suggested that the development of human language was actually the development of a communication technology that allowed a more efficient social organization. Suppose that a stable social organization depends on mutual reassurance among its members. This reassurance – whatever form it may take – consumes time. Dunbar hypothesized that in Chimpanzee societies, the mutual reassurance takes the form of a reciprocal service, namely grooming one another. Grooming is a very time-consuming activity. The limited number of Chimpanzees any individual Chimpanzee can regularly groom limits the size of stable societies of Chimpanzees to around 30. Humans do not groom each other (or at least not at the same scale), but use language to reassure each other. Based on his observations of Chimpanzee societies, Dunbar estimates the size of stable and dense human social groups to be around 150 individuals. If language, seen as a technology, is able to significantly expand the size of stable social groups that can be supported, then other technologies may have a similar effect. Riedl proposes that this may well be the case for social computing technologies. By letting us communicate more effectively, thus providing mutual reassurance, they may increase our potential to form groups. If our life is mirrored in real-time in a large online community – and we can see how the lives of others are mirrored as well – we in

fact live in a very large village. The global village then has its ultimate meaning: it is a village in which everybody is in tight social connection to everyone else, in which even small rumors spread quickly across the entire community, but in which the village has millions of inhabitants that live thousands of miles apart<sup>48</sup>.

It has been stated that apart from enabling communication and social interaction, social computing has a focus on social context. The social context of an individual user are the users she is connected to. In other words, the social context is the social network a user is embedded into. Social networks predate the age of computing, and an introduction into social network analysis will be given in the next section. It is important to note that social networks are a view of a collection of *social data*. We understand social data to be data about users and their interrelations. With the success of social computing applications, such data has become available in quantities that have previously been unthinkable. The evaluation and use of social data poses many challenges for various disciplines. In this part of the thesis, we would like to demonstrate how social data can be used to gain insight into social networks and their structure. By understanding collective social activity, we can access collective memory. Just as much as our individual documents and traces are part of our personal electronic memory, our social interactions are part of a social memory. The technologies of social computing define the way in which we shape this collective memory – it determines what we remember and what we forget, what we notice and what remains invisible, and ultimately, how our individual history is embedded into the history of our peers. An example for how social media shape collective memories is provided by Safrin and Schmidt [99]. Their *Pastiche* tool uses Twitter messages (tweets) related to the neighborhoods of New York City to generate a map of neighborhoods and associated topics. Figure 33 is a screenshot of their tool. It shows what people have been looking for in certain places, in a sense, it shows the history of places through the people that have been or lived there. Another interesting example, focusing on emotions and collective feelings, is *Wefeelfine* by Kamvar and Harris [100]. They search Twitter for tweets possibly containing emotions and aggregate these in several views, such as the *Mob* view, in which the amount of messages pertaining to a certain emotional state is displayed and can be further explored. Another view called *Metrics* tries to give an overview of the *state of the community* – in Figure 34, we can see that compared to earlier times, the observed community is feeling rather sad. While these are just early examples of the use of social data for extracting collective memories, they show the potential still awaiting exploitation.

---

<sup>48</sup>The only problem with the village metaphor is that one can escape from a real village and seek a good life elsewhere – this is hard in the case of the global village.



Figure 33: Tweet-based map showing different neighborhoods of New York City with associated topics. In the screenshot above, Chinatown is selected, the three associated topics are *prada*, *purses* and *coach* (Image: C. M. Schmidt [99])

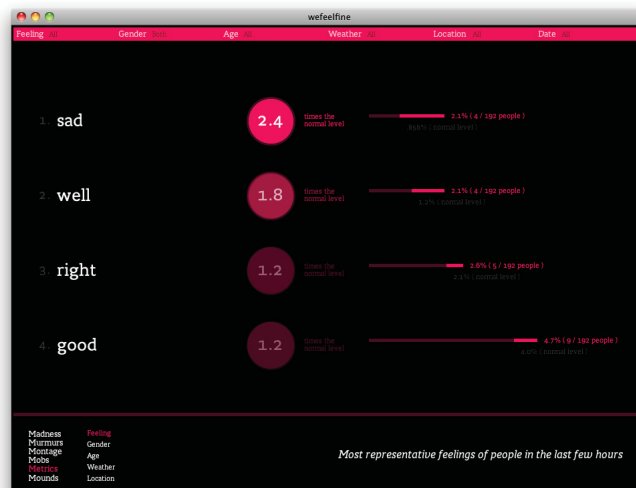


Figure 34: Overview of the emotional state of a community, based on previous measurements (Image: F. Müller, using [100])

The remainder of this part is structured as follows. We will first give an introduction to social network analysis and graph theory in section 3.2. Then, we introduce the social data we use, which has the form of a collection of email communications of employees of a corporation. In the remainder, we will show how this data can be visualized and analyzed to yield useful information, and describe the system architecture and implementation details of the application.



## 3.2 Social Network Analysis: Computations on Graphs

A social network is a network of *agents* for which *relational data* is available. The agents are typically individuals or organizations such as enterprises. The relational data describes how the agents are connected. It can in principle express any form of relation – mutual acquaintance, friendship, trust, cooperation, hostility, dependency, and so forth. The analysis of social networks is concerned with uncovering information that is implicit in the structural properties of such relational networks. In section 3.2.1, we introduce basic concepts of graph theory, which is an invaluable tool in social network analysis. In section 3.2.2, we describe questions that can be addressed to social networks, as well as related work on the subject.

### 3.2.1 Graph Theory

A social network consisting of agents and their relationships can be represented as a graph. A graph  $G = (V, E)$  consists of a set of vertices (also called nodes)  $V = \{1, \dots, n\}$ ,  $n \geq 1$ , and a set of edges,  $E \subseteq V \times V$ . An edge is an element connecting two vertices, and we will use the notation  $e_{ij}$  to denote an edge connecting the two vertices  $v_i, v_j$ . Two vertices connected by an edge are neighbors, and the set of vertices to which a vertex is connected is called its *neighborhood*. In an *undirected* graph, we have  $e_{ij} = e_{ji}$  for any  $1 \leq i < j \leq n$ , and in a *directed* graph, each edge has a directionality, such that  $e_{ij} \neq e_{ji}$  for any  $1 \leq i < j \leq n$ . A directed graph is also called *digraph*. A graph is *weighted* if either the vertices or the edges or both are associated with a label (e.g. a real-valued number), *unweighted* otherwise. In the case of social networks, we are working with directed or undirected, and possibly weighted graphs. We call a graph that represents a social network a *social graph*. Its vertices are the agents of the network, and its edges are their relationships.

A graph can be visually represented by a *graph drawing*, it should be noted, however, that any specific drawing of a graph is not constitutive for the graph. A graph is fully described by its adjacency matrix. An adjacency matrix  $M$  is a square matrix which has a row and a column for every vertex of the graph. The entry  $M_{ij}$  describes the connection between the two vertices  $i$  and  $j$ . In social network analysis, the adjacency matrix is sometimes called *case-by-case matrix* [101]. Figure 35 shows an adjacency matrix of a directed, unweighted graph with its drawing, Figure 36 shows an adjacency matrix for a directed, weighted graph with its drawing.

	1	2	3	4
1	0	0	0	1
2	0	0	0	0
3	1	1	0	1
4	1	0	1	0

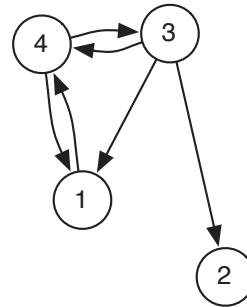


Figure 35: An unweighted, directed graph with  $V = \{1, 2, 3, 4\}$  and  $E = \{e_{14}, e_{31}, e_{32}, e_{34}, e_{41}, e_{43}\}$  given by its adjacency matrix (left), drawn on the right (Image: F. Müller)

	1	2	3	4
1	0	0	0	2
2	0	0	0	0
3	1	3	0	1
4	1	0	2	0

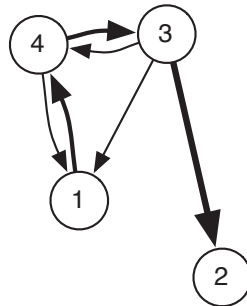


Figure 36: A weighted, directed graph with  $V = \{1, 2, 3, 4\}$  and  $E = \{e_{14}, e_{31}, e_{32}, e_{34}, e_{41}, e_{43}\}$  given by its adjacency matrix (left), drawn on the right. The weights of the edges are indicated by the width of the arrows (Image: F. Müller)

**Basic Properties of Graphs** A vertex has an associated degree  $d$ , which is the number of edges that are connected to it. In the case of a digraph, we have an in-degree, which is the number of edges for which the vertex is the endpoint, and an out-degree, which is the number of edges for which the vertex is the starting point. The degree can easily be read from an adjacency matrix: in an undirected graph, it is the number of non-zero entries in the row *or* the column; in a directed graph, the in-degree is the number of non-zero entries in the row and the out-degree is the number of non-zero entries in the column. The maximum degree  $d_{max}$  of a node is  $n - 1$ . The relative degree of a vertex is its degree (or in- or out-degree) in relation to the maximum possible degree,  $\frac{d}{d_{max}}$ . The *density* of a graph is given by the total number of edges in relation to the maximum possible number of edges. The maximum number of edges is  $\frac{n(n-1)}{2}$  in an undirected graph,  $n(n - 1)$  in a directed graph. The density  $E$  is a

measure for the overall connectedness of a graph. Figure 37 shows illustrations of various graph densities.

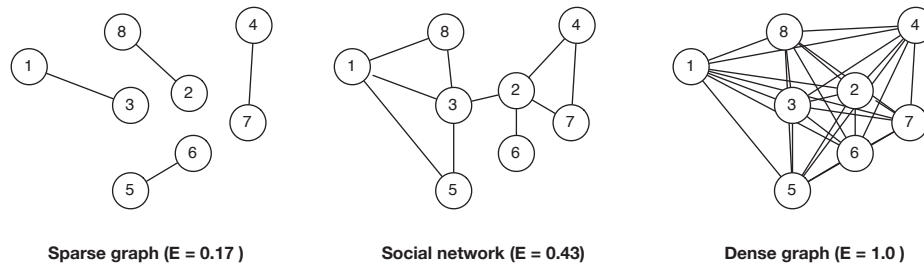


Figure 37: Unweighted, undirected graphs of varying density  $E$ . Preiss [102] proposes the following formal definition of sparse and dense graphs. A sparse graph is a graph in which  $|E| = \Theta(|V|)$ , and a dense graph is a graph in which  $|E| = \Theta(|V|^2)$  (Image: F. Müller)

If two vertices  $v_i, v_j$  are not neighbors, they may still be connected by a sequence of edges via other vertices. Such a sequence of edges is called a *walk* in the graph. If no vertex or edge occurs twice in a walk, it is called a *path*. Two vertices may be connected via multiple paths, of which the shortest is called the *shortest path*. There may be several shortest paths between any two vertices. The length of the shortest path between two vertices is the measure for the *distance* between them. In an unweighted graph (where each edge has a unit weight of 1), we can say that the neighborhood of a vertex is the set of vertices at distance 1. Accordingly, we can speak of a  $k$ -neighborhood, which is the set of vertices at distance less or equal to  $k$ . Note that this notion of distance applies to unweighted graphs. In weighted graph, edges may be assigned a label such as a number, and this label may be used as an indication of distance between the two adjacent vertices. If starting from any vertex, any other vertex in the graph is reachable via a path, the graph is connected. Otherwise, a graph is said to be disconnected.

**Centrality and Centralization** The centrality measure of a vertex aims at expressing the structural centrality of a vertex within a graph. In a social network context, this structural centrality is sometimes interpreted as importance or popularity ([101] p. 82). While centrality applies to individual vertices, centralization applies to the entire graph. A graph is considered centralized if there are a few vertices with a very high centrality, and it is considered decentralized if centrality is divided equally among the vertices. This concept is illustrated in Figure 38. Several measures for centrality exist. Among the ones we will discuss are degree centrality, closeness centrality and betweenness centrality. For a simplified treatment, we will always assume an undirected graph.

*Degree centrality* is the relative degree of a vertex, which is given by  $C_d(v) = \frac{d(v)}{(n-1)}$ . Since the degree only considers neighbors, it is a very local measure. It

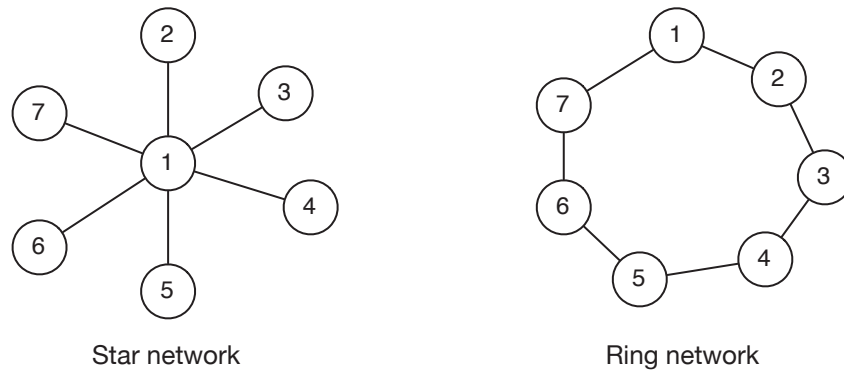


Figure 38: Illustration of network centralization. The star network is maximally centralized, with vertex  $v_1$  having a very high centrality, and all other vertices having minimal centrality. The ring network is maximally decentralized, with all vertices having equal (and equally low) centrality (Image: F. Müller)

can be made more global by including a certain  $k$ -neighborhood in the degree calculation. Depending on the density of the graph, however, this relaxation is only feasible up to a certain point. In a dense graph, every vertex will be in the  $k$ -neighborhood of every other node for a relatively low value of  $k$  (3 or 4 [101]). If in such a graph, degree centrality considers a far neighborhood, the measure loses its discriminative value.

*Closeness centrality* gives a measure of how far a vertex is from all the other vertices in the graph. A method for determining the closeness of  $v_i$  would be to sum the distances to all other points (see Equation 6). If we interpret distances in the graph as a cost of travel, for example, the vertex with the lowest sum of distances is the one that can reach every vertex at the lowest cost. In comparison to the degree measure, closeness considers the global context for every vertex.

$$C_C(v_i) = \sum_{v_j \in V} |v_i - v_j| \tag{6}$$

*Betweenness centrality* gives a measure of the importance of a vertex as a waypoint in a graph. The betweenness of a vertex  $v_k$  is understood in relation to a pair of other vertices  $v_i, v_j$  ( $i \neq j \neq k$ ). It is defined as the fraction of shortest paths between  $v_i, v_j$  that  $v_k$  lies on. If it lies on no shortest path between  $v_i, v_j$ , it has betweenness 0. If it lies on all shortest paths between  $v_i, v_j$ , it has betweenness 1. Suppose now we are interested of the betweenness of our vertex  $v_k$  regarding vertex  $v_i$ . In order to determine this, we must evaluate the betweenness of  $v_k$  for every pair  $v_i, v_x$  for  $x \in V, x \neq k \neq i$ . We sum the betweenness for every such pair, and we have the betweenness of  $v_k$  regarding  $v_i$ . We calculate this betweenness for all possible pairs  $v_k, v_i$ . The betweenness

of a point in relation to the entire network can then be derived as the sum of the betweenness values regarding all the other vertices:

$$C_B(v_k) = \sum_{i,j \in V, i \neq j} \frac{n_{ij}(k)}{n_{ij}} \quad (7)$$

where  $n_{ij}$  is the number of shortest paths connecting  $i$  and  $j$ , and  $n_{ij}(k)$  is the number of shortest paths connecting  $i$  and  $j$  that  $k$  lies on.

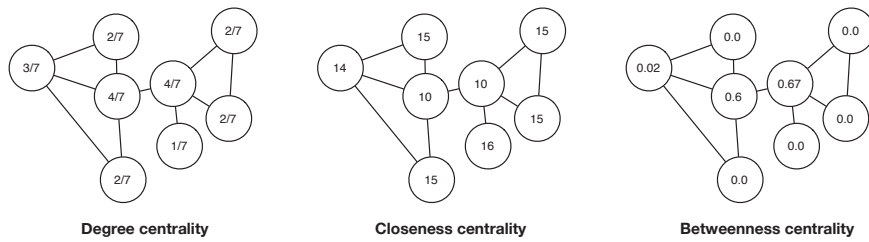


Figure 39: Comparison of various centrality measures in a unweighted, undirected graph (Image: F. Müller)

**Components, Cliques, and Clusters** So far, we focused on metrics for either the entire graph (connectivity, centralization) or individual vertices (degree, closeness, betweenness). In between, we have groups of vertices smaller than the entire graph. In a social network context, what we are particularly interested in are social groups, i.e. groups of agents that share some common properties that define them as a (social) unit. In graph theory, any subset of the vertices of a graph together with the edges between them is called a *subgraph*. The problem of finding relevant social groups is the problem of selecting a subgraph according to the relevant graph-metric criteria (a random subgraph will not likely correspond to a relevant social group). There are various concepts that describe aspects of selecting appropriate subgraphs.

A *component* is defined as the maximal connected sub-graph. In a graph that is not disconnected, there is only one component (the graph itself). In a directed graph, we differentiate between weak and strong components. In weak components, two vertices are said to be connected in both directions for an edge of any directionality between them. In strong components, two vertices are only said to be connected if there is an edge in both directions between them.

A concept closely related to components is that of the *clique*. While there are several definitions of what a clique is, it seems to be widely held that a clique is a *maximal complete subgraph* ([101], p. 114). A maximal complete subgraph is a completely connected subgraph in which every vertex is connected to every other vertex, and which is not part of another completely connected subgraph (maximal). Since in real social networks, the criterion of complete connect-

edness is rarely encountered, there are several relaxed definitions of a clique. In the case of a directed graph, *weak* and *strong* cliques are distinguished. A weak clique is a clique in which the determination of whether two vertices are connected ignores the directionality of the edges. In a strong clique, the directionality is considered, and two vertices are only considered connected if they are connected in both directions. An *n-clique* is a clique in which two vertices will be considered connected if there is a path between them with length equal or less than  $n$  (the clique according to the first definition is also an  $n$ -clique with  $n = 1$ ).

A *k-plex* can be seen as an extension of the clique concept. Then, a *k-plex* is a clique which is not fully connected. The  $k$  parameter determines how much less than fully connected the subgraph can be – for  $k = 2$ , the subgraph is considered connected if any vertex is not connected to at most 2 other vertices and connected to the rest. An illustration of these various concepts is provided in Figure 40. In any real social network, the application of cliques and  $k$ -plexes will produce overlap – points may be in many cliques or  $k$ -plexes, depending on the criteria that are applied for them. Alba ([101], p. 119) has formalized the idea of a *circle*. A circle would correspond to some higher social organization, and two or more cliques could be considered to form a circle if they have a certain proportion of overlapping members. Alba has proposed to merge cliques into one circle if two thirds of their members overlap (he would first identify 1-cliques of size 3, then merge them into a circle if 2 of the 3 members overlapped). In a second step, cliques would be joined to the circle with a relaxed overlap criterion (e.g. one third).

An important concept, which we will use later in this work, is that of a *graph clustering*. A graph clustering divides a graph into a set of *clusters*. Intuitively, a cluster is an area in which there is a high concentration of a certain element or property. We speak of clusters of cities, clusters of high-tech companies, and clusters of stars. According to Scott, clusters are defined by (a) their contiguity in the  $n$ -dimensional space where they are located and (b) their separation from other clusters ([101], p. 127). An illustration of this definition is given in Figure 41. While clustering is usually done for attribute data, it can be applied to relational data (see also Scott, *ibid.*). In principle, clusters can be identified either top-down (divisive) or bottom-up (agglomerative). In a bottom-up approach, we start with one or several individual vertices. We then try to combine it with other, similar vertices, to form clusters. In a top-down approach, we start with the graph as a whole and try to identify areas where groups of vertices can be separated, usually by an operation as cheap as possible, such as a minimal cut. A cut is an operation which divides the graph into two subgraphs, and it is considered minimal if no other cut exists that would result in the cutting of fewer edges.

What all clustering methods have in common is that they use a *distance metric* to measure distances between the individual components that are to be clustered. In order to have a workable example, we will try to cluster the

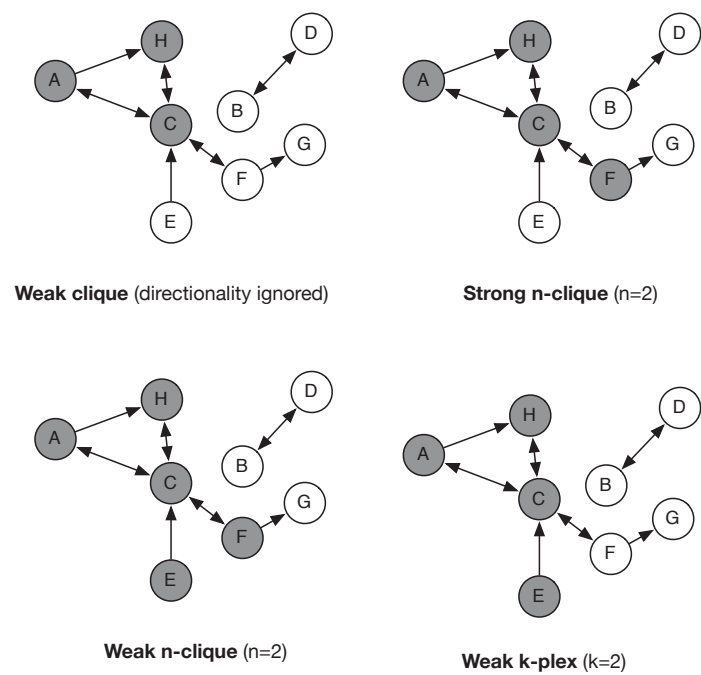
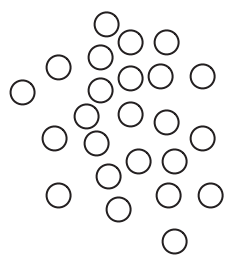
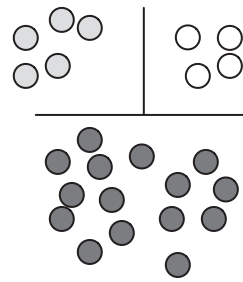


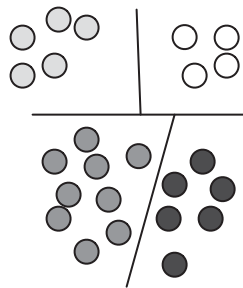
Figure 40: Groupings in an unweighted, directed graph, illustrating the differences between weak and strong cliques, n-cliques and k-plexes (Image: F. Müller)



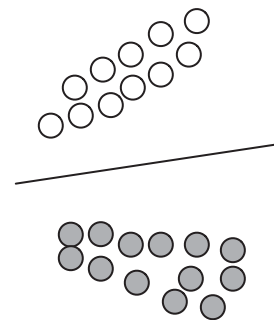
1 cluster



3 clusters (debatable)



4 clusters (debatable)



2 clear clusters

Figure 41: Data with various degrees of intuitive clusters (illustration purposes only, Image: F. Müller)



numbers  $\{1, 3, 12, 45, 47, 98\}$ . As a distance measure between the numbers, we just take their absolute difference  $dist(a, b) = |a - b|$ . Figure 42 illustrates two forms of clustering, *agglomerative* clustering on the left, *k-means* clustering on the right.

In agglomerative clustering, we start with  $n$  clusters (each element being a cluster). A distance threshold is defined, starting at 1. If for any distance threshold, the distance between two numbers is less or equal to the threshold, they are combined (agglomerated) into a new cluster, which becomes a number itself. The cluster has the numerical value of the sum of its constituents. The distance threshold is increased until all numbers are agglomerated in one single cluster. The result is a hierarchical clustering into one single cluster.

In *k-means* clustering, we start with  $k$  clusters, with  $1 \leq k \leq n$ , which are randomly selected among the numbers. Each cluster has a *location* (in a geographical interpretation, a center), which initially is the numerical value of the number which is selected. It is used to measure the distance between the cluster and any number. Every number is assigned to the cluster to which it has the least distance. Then, a new location for every cluster is calculated, given by the arithmetic mean of the numbers assigned to it. The two steps of assignment and re-calculation of the mean are repeated until the algorithm converges (i.e. the assignments, and in consequence the means, do not change any more). In the example, convergence is reached after three steps.

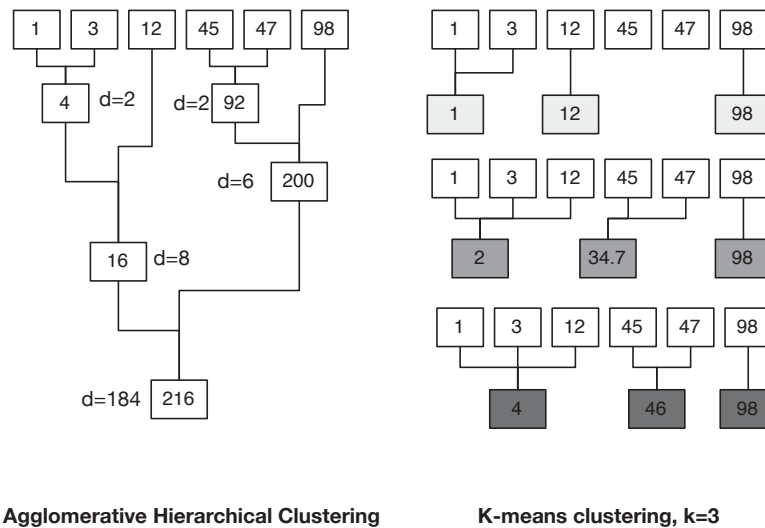


Figure 42: Illustration of Agglomerative and K-means clustering (Image: F. Müller)

If we interpret clusters in a graph as sets of vertices that share some common properties, the clustering of a social graph would result in the detection of relevant social groups. In sections 3.4 and 3.5, we will use agglomerative graph

clustering to derive the organizational structure of a professional social network. We will introduce a novel distance measure taken from the visualization of that social network.

### 3.2.2 Questions to Social Networks

Social network analysis is a field most strongly associated with the social sciences. Its core focus is the understanding of the behavior of agents by observing that behavior in the context of the behavior of other agents. This *networked view* is not confined to social networks. Network analysis is an approach encountered in astronomy, technical networks, genetics, and neurology (for a broad overview of the networked approach, see [103]). In social network analysis, the agents observed range from human beings to organizations, states, and social animals like bees or apes ([104], p. 2).

Three main traditions are usually identified as the predecessors and founders of social network analysis [101] [104]: the sociometric analysis pioneered by Jacob Levy Moreno; the analysis of interpersonal relations and the identification of social groups by researchers at Harvard university; and the *Manchester anthropologists* who used the former instruments to analyze structural properties of small societies. Moreno introduced the concept of the *sociogram* to formally represent social groups. By representing group members as points and their relationships as lines between the points, sociograms can help identify important structural aspects of groups (such as a sociometric star, with one individual in the center connected to all other individuals of the periphery, [101], p. 10). At Harvard, the technique of the sociogram was later used in first broad studies. An example are Elton Mayo's Hawthorne experiments. Between 1927 and 1932, Mayo investigated the productivity and work conditions at the Western Electric Hawthorne Works in Chicago. He gathered detailed data on 14 employees mostly via interviews, and used this data for his analysis. An important contribution was a sociogram of a specific organizational unit (the bank wiring room), which reflected the actual, informal organizational structure of the group. It contrasted with the official (formal) organizational chart, and thus proved that valuable insight into group structure and dynamics could be expected from social network analysis. At the university of Manchester, a first formal framework for social network analysis was developed in the 1950's. It attempted to unite a formal description of social networks (as, for example, sociograms) with the sociological interpretation of the evident structures and relationships. Mitchell, a late exponent of the Manchester tradition, contributed important initial work on finding rather simple measures for interpersonal relationships, such as reciprocity or durability ([101], p. 31). At Harvard university, a breakthrough in social network analysis was reached when formal mathematical methods were introduced into the study of social relations. Most notably, *graph theory* has proven to be a decisive and long-lasting contribution to the field ([101], p. 28).

The view of social networks as graphs, the separation of actual material properties of social networks and the focus on relational aspects allowed the application of complex computations in social network analysis. This development is best illustrated by the increasing size of the data sets.

For many years, the social networks investigated included only a small num-

ber of agents. In the Hawthorne experiments, the data set included 14 employees. Milgram’s famous small world experiment (which, today, is often interpreted as the *six degrees of separation* principle, i.e. the assumption that from any individual, there is a path through the social network to every other individual with a length of six edges [105]) involved a total of a few hundred participants, with a starting number of participants of 44. In 1977, Zachary published his analysis of the social network of a Karate club at a U.S. university, consisting of 34 members [106]. Social network data available for research today shows considerable differences. They are not manually collected through interviews, but rather automatically generated from large online (or offline) data sources. Table 5 lists several social data sets available from the —Large Network Collection of the Stanford Network Analysis Platform (SNAP [107]). The Enron set provided is based on the Carnegie-Mellon version (see Section 3.3.2). The Wikipedia edit history is an affiliation network with two types of nodes users and pages. The edges represent users editing pages.

Dataset	Nodes	Edges
Enron email (partial)	36,692	367,662
Trust network from Epinions.com	75,879	508,837
Slashdot social network November 2008	77,360	905,468
Amazon product co-purchasing 2/3/2003	262,111	1,234,877
LiveJournal online social network	4,847,571	68,993,773
English Wikipedia edit history	5,800,800	250,000,000
Twitter June-Dec2009 collection	17,069,982	476,553,560

Table 5: Selection of social data available from the Stanford Large Network Collection (Source: [107])

As evident from Table 5, social data with millions of nodes and interrelation are not exceptional. And the possibility for even larger corpora is demonstrated by the number of users that engage in frequent online social interaction. According to a collection on Wikipedia [108], there are over 10 social computing platforms (*virtual communities*) with more than one hundred million users. In its usage statistics, Facebook announces [109] that as of March 2011, there are 500 million active users, of which half log on to Facebook every day. Average users create 90 pieces of content (ranging from pages to messages and pictures) every month, and more than 30 billion pieces of content are shared each month. Evidently, this amount of social data goes far beyond the data sets that are currently readily available for research. It can thus be concluded that in the future, the size of social data corpora will increase further. At the same time, the aspects of the real world that are covered by this social data will continue to expand.

By viewing social networks as social graphs and using graph theoretic methods to analyze them, they become comparable among one another and over

time. Several important graph metrics have been introduced in that regard (see Section 3.2.1). Overall properties such as density, centralization or connectedness can be used to compare different networks. Vertex metrics such as degree, closeness or betweenness centrality can be used to identify the role (or importance) of individual nodes within networks. The formation of groups (cliques) and the clustering of graphs help in identifying organizational properties of the network. In short, all the interesting properties assigned to graphs in graph theory can potentially be applied to social networks. How they relate to them, i.e. what their meaning in a social context is, is a guiding motivation for researchers of many fields.

There are many specific applications for which social network analysis is considered to be an appropriate method. In epidemiology, social network analysis is used to model the spreading of diseases. Eubank et al. [110] have recently proposed a method for the early detection of disease outbreaks in urban environments. Valente [111] has applied social network analysis to the investigation of *innovation diffusion*. Several approaches exist for predicting physiological and psychological health based on individual's social networks [112] [113]. In security applications, social network analysis is used to detect potential threats [114]. While the (incomplete) range of topics is broad, insights into the relationships of the agents of the network are, in almost all cases, based on common techniques. This is exemplified by Hansen et al. [115], who have analyzed a mailing list in order to identify potential candidates for the position of future list administrators. Based on a graph representation of the mailing list, they find vertex properties characteristic of administrators, and then search for non-administrator vertices that have similar characteristics. In addition, they can simulate the removal of administrators and the effect this (and a subsequent choice of an administrator) has on the community structure.

With social network analysis, general structural properties of social networks can be analyzed. Individual nodes, as well as partial communities or entire networks are made comparable and can be characterized in socially relevant terms. Based on this, specific questions to specific networks can be answered. These questions can concern the current state of the network or individual vertices as well as the prediction of future development. In the remainder of this part, we will apply social network analysis to a social network present in the form of an Email corpus. We will focus on the detection of community structures.

### 3.3 Introduction to the Enron Email Dataset

This section introduces the Enron Email dataset, which is also referred to as the *Enron corpus* or the *Enron dataset*. The Enron corpus is a collection of emails from Enron employees which has been published in the course of the *Western Energy Crisis* investigation by the United States Federal Energy Regulatory Commission (FERC) in 2003 [116] and is now in the Public Domain. The original dataset consists of over a million individual messages contained in a total of approximately 150 mailbox directories (see Section 3.3.2).

Since its initial publication by the FERC, the corpus has been used by various researchers. Multiple inconsistencies and errors have been found, and several *cleansed* versions of the corpus have been derived and published for use by other researchers (see [117] for a collection of various data sets). The cleansed versions of the corpus are not identical, since different cleansing strategies were employed. However, it has been established that the Enron corpus (in any version) contains around a quarter of a million unique emails. This number is exceptional, since no other such large collection of corporate Emails is known to be publicly available. While today, dedicated social network sites such as Facebook have far greater collections of social data, they are not as openly accessible as the Enron dataset. In this regard, we consider the Enron dataset to be one of the most interesting sets of social data currently available for analysis.

This section is structured as follows. First, we briefly recapitulate the history of the Enron corporation. We then examine the context in which the Enron corpus was published, and look at its structure. Finally, we describe related work that has been conducted on the corpus since its publication.

#### 3.3.1 A Brief History of Enron

In 1985, Enron was founded through the merger of Houston Natural Gas and Internorth, an energy company from Omaha, Nebraska. Its initial business was the operation of a network of gas pipelines in the United States. The natural gas business was characterized by very strong regulations and very little market dynamics. Both producers of natural gas and the buyers had long-term, fixed-price contracts with distribution companies such as Enron. There were, however, attempts to deregulate the natural gas market, and as first deregulation measures took place, the business of Enron changed. It was now possible to trade gas on the spot, i.e. to dynamically fix a price between seller and buyer. Enron started to use spot trading on the producer side – i.e. they would still have long-term fixed-price delivery contracts with energy users, but would buy the gas on the spot market. Enron would thus take the risk of rising gas prices, and speculate on gas prices being stable or even sinking. Enron soon expanded its business to energy trading, and over the years, its trading portfolio would include chemicals, pulp and paper, steel, water, and even weather risks.

From on the early nineteen-nineties, Enron followed an aggressive expansion strategy. In order to raise more capital for acquisitions, it resorted to bookkeeping practices that were not previously known in the energy industry. First, it used mark-to-market accounting, which allowed it to disclose future earnings as current profits. If it entered a 10-year delivery contract with a buyer of gas or electricity, it was able to book the present value of the entire (future) contract at present time. Second, it developed techniques to keep losses and debts off its balance sheet by integrating them into supposedly independent special purpose vehicles. While these practices were very uncommon and novel for the energy industry, Enron's auditing firm, Arthur Andersen, gave green light to all of the accounting measures later recognized as illegal.

The fast expansion of Enron made it one of the most successful companies of the nineteen-nineties. From around \$10 in 1990, the stock price rose to \$40 in 1998, and \$89 in September 2000. Enron reached a peak market capitalization of over \$60 billion, and employed over 20'000 people. Having been termed "America's Most Innovative Company" by the Forbes magazine four years in a row [118], Enron files for bankruptcy on December 2nd, 2001, with its stock being valued at \$0.6 a share. The deep fall of Enron had a significant impact. In the aftermath of the bankruptcy, several key employees of Enron were found guilty of various economic crimes and received sentences of up to 24 years in prison. Arthur Andersen, once one of the most renowned auditing companies of the world, was found guilty of criminal charges in relation to its auditing activities at Enron, and has discontinued its auditing business in 2002. A large number of Enron employees, who had invested their pension funds in company stock, have lost most of their pension savings. The developments around that time are evident in the development of the share price, depicted in Figure 43.

The prosecution of illegal bookkeeping practices of Enron was only one of two major official investigation into Enron's business. The other one was concerned with possible market manipulations through energy traders in the *Western U.S. Energy Crisis* from 2000 to 2001. In 1996, California started a process of partial deregulation of its energy market. A first major step was the introduction of a spot market for electricity, which began operating in 1998. The deregulated energy market was a very complicated construct. In summer 2000, energy prices in the Western U.S. increased sharply, accompanied by repeated rolling black-outs. A rolling black-out is a deliberate measure taken by the electricity grid management to prevent a total black-out, i.e. a total loss of power over the entire network. In a rolling black-out, parts of the grid are periodically taken off-line (they lose power), such that at any given time, a part of the grid doesn't have power. This way, the overall load of the grid is lower. The circumstances in which rolling black-outs are required are when either, not enough electricity is produced to meet the demand throughout the grid, or, if enough electricity would be available, but the transportation capacities of the grid are not sufficient to transfer the power to where it is needed. It was later found that Enron, as owner of several power plants, had deliberately taken some of them off-line to

Enron Share Price Closing 1997-2002

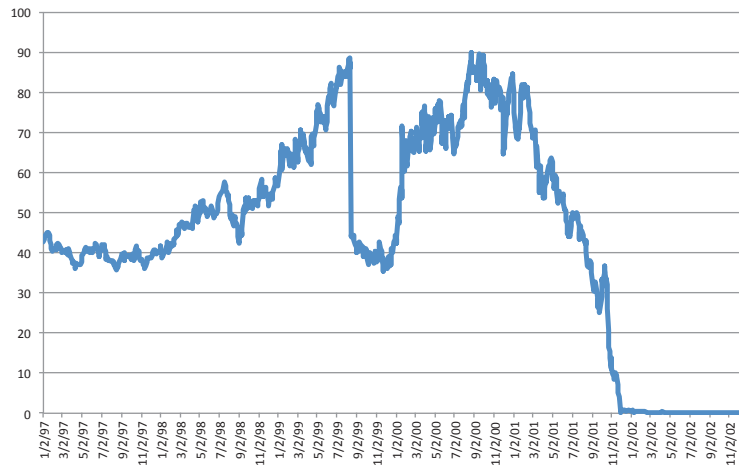


Figure 43: Development of Enron Stock Prices 1997-2002 (Image: F. Müller, Data: [119])

produce a shortage. This would not only lead to such rolling black-outs, but also to much higher electricity prices. For example, in California, the average price for one megawatt-hour in December 1999 was \$45, while one year later, in the midst of the crisis, it was \$1400. In 2001, the FERC ordered a 'calming' of the Western energy market by ordering generators to run and by defining price caps. Enron had designed several special strategies in order to maximize its profit in the market environment of California after the partial deregulation. Among them was one called *Ricochet*, which also came to be known as "megawatt laundering" [120]. Enron would buy electricity in California and then export it to buyers in neighboring U.S. states. It would then buy it back (for a higher price) and import it into California. Thus, it was able to charge higher prices that it could have by selling the power directly in California, since state-internal pricing was eventually capped by regulatory measures. All in all, it is estimated that the Western energy crisis cost California somewhere between \$30 and \$45 billion.

In the course of the FERC investigation, Enron was ordered to release a large amount of information in electronic form. Among it are the mailboxes of several key employees that were thought to be relevant in the context of determining whether Enron was involved in wrongdoing. In addition, many scanned documents, audio files of telephone conversations, and other relevant material have been published<sup>49</sup>. Before we enter into the details of the Enron

<sup>49</sup>The data is still available on the FERC website, together with additional explanatory materials and reports [116]



corpus, it should be noted that the publication of the emails has received some criticism. The Wall Street Journal noted that Enron's employees did not deserve to have their email communications – in which there are many private messages regarding sometimes delicate issues – “put on public display” [121]. Several messages have been removed from the corpus, either before publication on Enron's request, or after initial publication at the request of affected individuals. Regardless of whether the publication of these Emails was justified or not, they should be treated with care for the privacy of the individuals they belong to. The author tries to do his best to protect the privacy of the individuals involved.

### 3.3.2 Properties of the Dataset

The exact number of emails as well as the number of users of which the mailboxes were collected from are debated. Krasnow [122] suggests that the FERC has published approximately 1.4 million emails, of which many were empty or duplicates. These issues were addressed by several groups who have worked on the Enron corpus and created a cleansed version of it. In the process of cleansing, they removed duplicate or empty emails, identified inconsistencies with the users mailboxes, and collapsed several email addresses to one single email address (e.g. any non-existing address of the form `x@enron.com` could be replaced by `invalid@enron.com`). According to William Cohen of Carnegie Mellon University, the Enron email dataset was initially purchased by Leslie Kaelbing of the MIT after it had been published by the FERC [123]. The data was then edited by the SRI research institute, and made publicly available free of charge. Carnegie Mellon offers this raw version of the corpus (with many duplicate emails), together with background information on the corpus (see *ibid.*). This raw version seems to be the basis of most other versions of the corpus. All the attachments of the emails, which are usually not included in the corpus, can be obtained from the Electronic Discovery Reference Model (EDRM) [124].

Regarding the number of users from which the emails have been collected, the conclusions of the various groups working on the corpus differ. In the FERC version, the Email data was collected from 158 individual users' workstations. However, it is possible that two users represent one and the same person (e.g. Stephanie Panus existed as a user *phanis-s* and *panus-s*, so a misspelling, and for Lawrence Whalley, there existed two users, *whalley-l* and *whalley-g*). It has also been mentioned that the emails for one user only consist of automated calendar notifications [125]. Table 6 lists several versions of the corpus with the number of emails present and the number of *custodians* (i.e. individual users). As can be seen, the latter number lies between 148 and 158.

Dataset	Custodians	Emails
FERC	158	1,400,000
CALO	151	517,431
Shetty and Adibi	151	252,759
Carreda	147	250,484

Table 6: Various versions of the Enron corpus. CALO stands for *Cognitive Assistant that Learns and Organizes*, a Carnegie Mellon project. The Shetty and Adibi corpus was developed at the University of Southern California [126]. The Carreda corpus was developed at the University of Massachusetts [125]. The EnronData.org corpus is available at [117].

Apart from the question about the number of custodians, there is an issue with the number of email addresses that exist for each custodian. In most cases, it is not a one to one mapping. For example, Vince J. Kaminski has several

email addresses, such as *j.kaminski@enron.com*, *vince.kaminski@enron.com*, or *kaminski@enron.com*. In order to assign these several addresses to the same person, no definitive solution has been found. Diesner et al. [127] have used text matching techniques and have increased the address-to-individual ratio from 1.0 (in the original data set) to 2.2. Our approach, which consists of a semi-automated expansion of the mapping, will be described in section 3.5.2.

In this thesis, the dataset provided by Shetty and Adibi at the University of Southern California is used [126]. There are 151 individual users and 252,759 emails. The cleansing that Shetty and Adibi performed consisted of the following steps. To eliminate duplicate messages, they followed a folder-based approach, identifying email folders that contained duplicates of emails already present in other folders. As the authors point out, many of these duplicates-only folders were automatically generated by applications. All such folders were removed from the collection. Then, all messages which were considered to contain 'junk' data, such as lines of text originating from past attachments, were removed. Third, email addresses considered to be invalid were changed to *no.address@enron.com*, and all messages for which no recipients were disclosed (using different labels) were changed to *undisclosed-recipients@enron.com*. Finally, messages returned by a mail server (e.g. delivery failure) were removed. The corpus by Shetty and Adibi is provided as a MySQL database dump, details on database organization and message retrieval are given in section 3.5.2.

The custodians are the individuals for which the 'complete' email exists – i.e., their mailboxes (either locally on their computer, or on their server account) have been completely taken over by the FERC. This does not mean, however, that there does not exist considerable email traffic for other individuals at Enron. Table 7 lists several individuals together with the number of emails that have been found for them in the database. The top half of the table lists the 10 custodians with the most messages, while the bottom half lists the 10 non-custodians with the most messages. It can be seen that, for example, there are more messages for Mark Taylor, whose complete email traffic is not known, than for 147 of the custodians. These individuals provide an interesting hypothesis space: we know much, but not all about them. However, as will be detailed later, the focus of the analysis lies on the 151 custodians, for which we have the most complete information. At least one investigation has chosen to include this partial information, building a corpus that includes non-custodians and resulting in a total of over 600 individual users [127].

The total of 252,759 messages in the Shetty and Adibi corpus date from roughly four years. Some timestamps of the messages are obviously invalid (e.g. '0001-05-30 13:10:06' or '2044-01-04 14:48:58'). After eliminating them, the messages date from October 1998 to December 2002. It should be noted that the number of messages is not evenly distributed in time. Only very few messages are available for 1998 and 1999, as well as for 2002. Most messages are from the years 2000 and 2001, with a peak count of 29,556 messages per month in October 2001. Figure 44 shows the number of emails per month for the years

<b>Name</b>	<b>Total Emails</b>	<b>Sent</b>	<b>Received</b>
Jeff Dasovich	16,155	6,285	9,870
Sara Shackleton	12,962	4,810	8,152
Tana Jones	12,866	4,437	8,429
Richard Shapiro	9,652	654	8,998
Kay Mann	8,439	5,100	3,339
Louise Kitchen	8,262	1,510	6,752
Gerald Nemec	7,131	2,275	4,856
Chris Germany	6,229	3,686	2,543
Sally Beck	5,547	1,596	3,951
Elizabeth Sager	5,508	1,524	3,984
Vince Kaminski	10,265	5,843	4,422
Mark Taylor	8,870	2,081	6,789
Pete Davis	7,552	2,500	5,052
Steven Kean	7,040	1,796	5,244
John Lavorato	6,710	1,525	5,185
Susan Mara	6,065	847	5,218
Paul Kaufmann	4,888	215	4,673
Tim Belden	4,501	286	4,215
Richard Sanders	4,220	1,639	2,581
Harry Kingerski	3,685	143	3,542

*Table 7:* Number of emails for top 10 custodians (top) and non-custodians (bottom). As can be seen, the top non-custodians are very active in communication, however, since not all of their messages are known, they are not included in the key focus group (Source: Shetty and Adibi dataset)

1999-2002. The low numbers for 1999 and 2000 do not indicate that only few emails have been sent in these years, rather, it is the result of several factors, such as limitations in mailbox space (resulting in old emails being deleted to free up space) and email retention policies (automatic deletion of email that have reached a certain age) - the older emails are, the more likely they are deleted. Nevertheless, the development of the total number of emails is an indication of what happened at any given time in some cases. Holiday seasons (Christmas and the summer) can be identified by the short depressions, indicated in the figure. The steep rise in October 2001 happened when the crisis reached broad publicity and escalated. Evidently, Enron employees have communicated most often at that time of extreme change. Also, the steep decline in December and early 2002 can be explained by the crisis having gone past its climax, another holiday season, and the massive layoffs that followed after Enron had filed for bankruptcy.

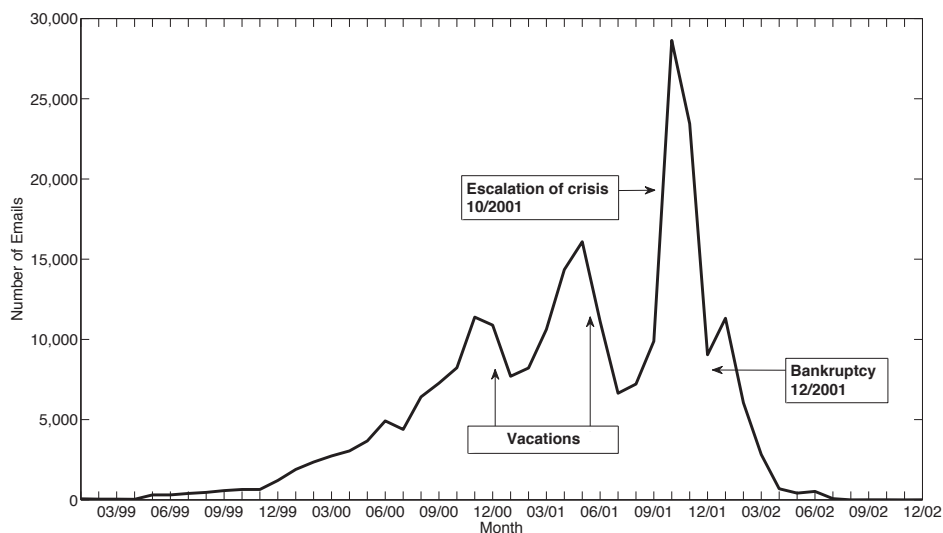


Figure 44: Distribution of number of emails over time, in a month-wise sampling. There is a peak in October 2001, the time when the Enron crisis was at its most intense moment and evolved into the Enron scandal (Image: F. Müller, data: Shetty and Adibi dataset)

Table 8 shows the number of unique email addresses for several email domains. The great majority of email addresses belong to the enron.com domain, which could be expected. The email addresses belonging to the domains aol.com, hotmail.com and yahoo.com are mostly private Email addresses of employees. Dynegy.com and duke-energy.com are domains belonging to Enron competitors. Haas.berkeley.edu is the domain of the Haas Business School (University of California, Berkeley). From the distribution of email addresses,

it seems that Enron-internal communication dominates the corpus, with private email addresses of employees having a nevertheless important role (some of their use is for private matters, some also for work-related matters).

Domain	Count	Description
enron.com	34,772	Enron Corporation
aol.com	3,266	Private Internet Service Provider
hotmail.com	1,985	Private Email provider
yahoo.com	1,427	Private Email provider
haas.berkeley.edu	697	Hass Business School
msn.com	461	Private Email provider
earthlink.net	447	Private Internet Service Provider
dynegy.com	350	Competitor
houston.rr.com	289	Private Internet Service Provider
worldnet.att.net	257	Private Internet Service Provider
duke-energy.com	251	Competitor

Table 8: Number of unique email addresses for the most prominent domains (Source: Shetty and Adibi dataset)

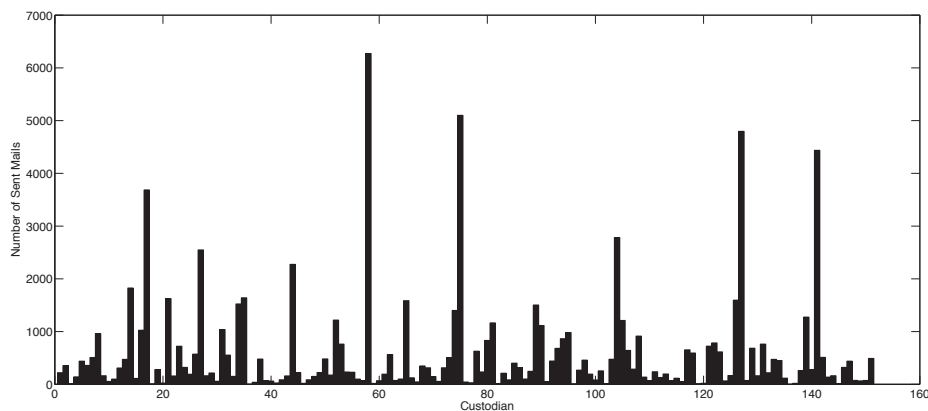


Figure 45: Number of sent emails of all custodians (Image: F. Müller, data: Shetty and Adibi dataset)

As evident from Table 45, the emails are not equally distributed over the custodians. The figure depicts the number of sent emails. Only a few custodians have over 4,000 messages, and most custodians are in the range of several hundred to one thousand messages. In an analysis of the distribution of the number of email messages over the hour of the day, we have found that it is subject to significant change over the months. Figure 46 shows this distribution

for the months of January 2001 and September 2001. While in January, the arithmetic mean lies around 6:30h in the morning, it lies around 11:00h in September. If we look at the mean over the entire year, we can see that it lies between 06:30h and 06:50h from January to March, between 8:45h and 9:15h in April and May, around 10:00h in June and around 11:00h from July to December (which means a shift of around 4, 2 or 1 hour). We have not come to a conclusive interpretation of this shift. Krasnow noted that there are inconsistencies in the time zone assignments between the original FERC and the UCB dataset ([122], p. 51). However, there is no mention of inconsistent time stamps within the same corpus. Further investigation is required to determine the cause for the observed shift.

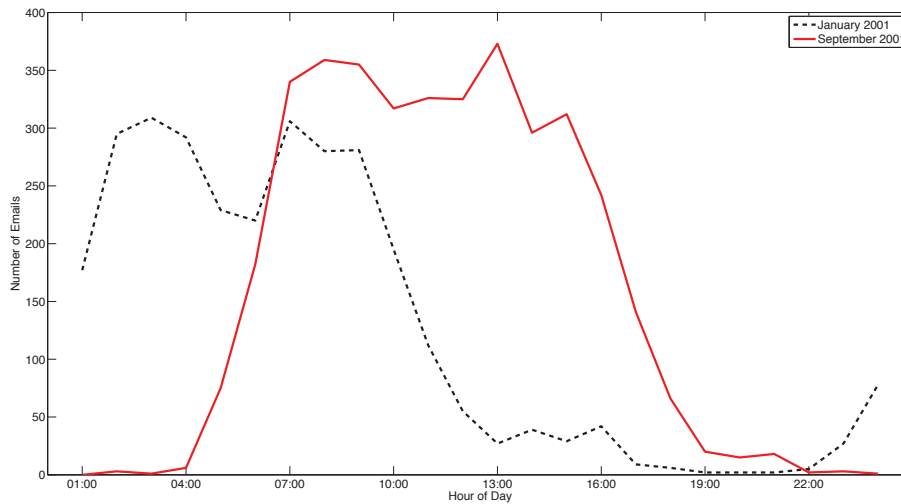


Figure 46: Distribution of the number of messages over the hour of day for the months January 2001 and September 2001 (Image: F. Müller, data: Shetty and Adibi dataset)

### 3.3.3 Related Work on the Enron Dataset

Since its publication, the Enron corpus has been used by various researchers. Initial work has focussed on the overall properties of the corpus and the elimination of duplicate and erroneous data. Later, two principal approaches can be distinguished. One focuses on the structural properties of the communication network and is based on graph theoretical methods, the other makes use of natural language processing and tries to analyze not only the structure, but also the content of the communications. Several groups have combined both approaches. An overview of these works is presented here.

Krasnow has reported on several aspects of the Enron corpus data. Among other things, he pointed to problems in the handling of time stamps. He found

inconsistencies for some messages between the original FERC dataset and the UCB dataset ([122], p. 51). He also details the changes made in various versions of the corpus, and includes a discussion about the benefits and risks of de-duplication (i.e. removing messages *considered* to be duplicates) of the corpus. A good overview of the problems associated with effectively cleaning the data set is provided by Zhou, especially concerning the collapsing of several addresses to one individual [128]. An annotated version of a subset of the corpus (January to December 2001) is provided by Berry et al. [129]. It contains approximately 5,000 emails, which have been manually indexed into 32 topics (e.g. legal issues of the Western Energy Crisis, emails about the downfall of Enron). It is well suited for an approach that combines structure *and* content of the corpus.

The questions researchers expect to find answers to in the Enron corpus range from very specific to rather general. Among the specific questions, several prediction tasks are notable. Bekkerman et al. [130] as well as Klimt et al. [131] have used the organization of emails into folders (manually conducted by the custodians) to try to automatically classify messages. Klimt et al. use a support vector machine to classify messages based on individual message headers (e.g. 'From', 'To') or on the content of the messages. Bekkerman et al. also work with support vector machines, they focus on the content of the messages (as Klimt et al. in *bag-of-words* representation). Nussbaum et al. [132] try to predict the importance of email messages based on data from the Enron corpus. For that purpose, they track the history of communication pairs, however, their results seem not to be conclusive.

Among the more general questions are questions about the qualification of social relations. While in social network analysis, relationships of trust have traditionally received broad attention, this is not the case for the Enron corpus. One reason might be that an important source of trust information in email communications, the *blind carbon copy*<sup>50</sup> field, is unavailable in the Enron corpus. It seems (and this applies across corpus versions) that somehow, the *carbon copy* and the *blind carbon copy* fields have been mixed up in the generation of the data. In the version of Shetty and Adibi, there are 253,735 carbon copy recipients and 253,713 blind carbon copy recipients, and they correlate. In other words, in almost all cases, whenever there is a CC recipient, that same recipient also has a BCC. We have not found a reason why in 23 cases, there is no corresponding BCC recipient for the CC recipient. Some authors have used the BCC field in their research, however, they do not account for the strange correlation [133] [134]. We consider that the potentially valuable information of sending blind carbon copies – indicating trust – has been compromised in the

---

<sup>50</sup>The recipient of a blind carbon copy (BCC) of an email is not visible to any other recipients of the message. In one possible interpretation, this can imply two things: first, that the trust the sender puts into the direct and non-blind copy recipients of the message is limited, and second, that the trust the sender puts into the blind carbon copy recipients is considerable (since the sender relies on the BCC recipient to not inform the other recipients about receiving a blind copy).



Enron corpus<sup>51</sup>. Instead of trust, the aspect of social hierarchy is a direction of investigation often encountered in work on the Enron corpus.

Diesner et al. [127] are interested in the dynamics of the relationships between different hierarchy levels. They have built a custom corpus on the basis of the corpus by Shetty and Adibi, in which a total of 676 individuals (they include individuals which occur in the custodians email often, but which are not custodians themselves) are associated with a position and a rank within the corporate hierarchy. The position refers to a job description, such as *CEO*, *manager* or *trader*, while the rank is associated with a hierarchy level, such as *board*, *executive management* or *senior management*. Based on the analysis of individuals, positions and ranks in terms of several metrics such as density and centrality, Diesner et al. try to infer general statements about the corporate communication network during the development of the Enron crisis. They have, for example, found that senior management is most active in terms of sending messages, that higher ranks effect more top-down communication than lower ranks (i.e. it can be assumed that they are more involved in the distribution and control of work), and that the most lateral communication (within the same rank) occurs in senior management. Creamer et al. [135] are interested in automatically detecting social hierarchy. They use the Shetty and Adibi corpus version, together with a subset provided by FERC that only contains the 54 custodians employed at the Enron North America West Power Trade desk. The authors introduce the concept of the *social score*, a scaled number between 0 and 100 that can be assigned to every individual in the network. It is a combination of 11 metrics which are calculated for every user. Among the metrics are the number of emails and the average response time, a clique score (related to the number and size of cliques an individual belongs to), and several measures of vertex centrality. A categorization of employees based on the social score showed promising results, both in the case of the Enron corpus and on the email communications of the university group the authors belong to.

As an exponent of investigations that include an analysis of email content, McCallum et al. [136] should be noted. They present an author-recipient-topic (ART) model for social network analysis. Instead of just tracking an author-recipient pair, they identify topics and associate them with author-recipient pairs (thus building an affiliation network). The authors suggest that mere connectivity information (i.e. structural information of the social network) is not sufficient for adequately identifying roles and functions within the network. In a statistical analysis of authors and author-recipient pairs in the Enron corpus, a power-law distribution as typical for social networks is substantiated. The authors show how they have identified 50 topics, given by groups of frequently co-occurring words and afterwards hand-labeled (e.g. *legal contracts*, *government relations*). When including the topics in the characterization of relationships, the authors claim they have found a better measure for determining role similarity (or dis-

---

<sup>51</sup>We have chosen to disregard the BCC field entirely

similarity). They apply their scheme not only to the Enron data, but also to the set of personal email communications of one of the authors. In the latter case, they have better knowledge of the exact roles and functions of the individuals involved, and can better verify their results. Compared to traditional social network metrics and a mere author-topic or author-recipient consideration, the ART model seems better able to identify similar and dissimilar roles within the network (e.g. 'ML researcher on SRI project', 'UMass admin assistants'). In a more recent study, Hossain et al. [137] investigate the relationship between the centrality measure of a network node and the amount of coordinative work they perform. They evaluate degree, closeness and betweenness centrality. In a first step, they calculate a coordination score for every individual using text mining techniques (i.e. by analyzing the content of the emails people have sent or received). These degree scores are used as a reference to test various centrality scores, which are calculated based on the structural properties of communication. The authors test several hypotheses, and propose that betweenness centrality is the centrality measure most accurately reflecting coordinative activity, and that when working with a directed graph, out-centrality correlates more to coordinative activity than in-centrality (which, intuitively, makes sense, since coordination is an activity requiring the distribution of information).

Finally, the Enron Data Reconstruction Project, initiated by John Wang, should be mentioned [117]. It provides a complete version of the corpus, including attachments, in cooperation with the EDRM (Electronic Discovery Reference Model) project. It aims at integrating the knowledge gained so far by the various groups that have offered a corpus version into one final version. An overview of other versions of the corpus is provided – currently, the list links to 10 versions. Also, a description of a wide range of research projects related to the Enron corpus is provided. It seems that although the Enron corpus predates the rise of social networking platforms and can be considered a rather small corpus compared to current trends in social data, it remains a popular research object. This, in part, may be due to the unspecific nature of email communication. Opposed to social networks, where social relations and attitudes are expressed explicitly (in *friendship*, *likes* and *dislikes*) and attributive data is precisely stored (such as age, gender, professional title, religious and political views), the quality of the relationships evident in email communications must be *inferred*. The generality of the content of the communication leaves the field open for a broad range of research questions. In addition, since the corpus has not been anonymized, and a lot of organizational history is known about the Enron corporation, the results of the work performed on the data can be tested in detail.

### 3.4 A New Approach to Sampling Social Graphs

This section introduces the approach we have developed to analyze social graphs. We focus on the problem of identifying relevant clusters, i.e. we are interested in social groups at various levels. We do not confine ourselves to a specific type of group (such as *close work group*), but want to determine the overall organizational structure, starting with small groups and eventually spanning the entire organization. We consider two aspects to be of special importance.

First, the temporal development of the social network must be taken into account. When constructing a graph for a social network, the activities in the network are usually sampled over a certain period of time. The graph represents the accumulation of these activities, and does not account for their development over the time period. The larger the time period becomes, the more information is lost in the accumulation approach. We propose a technique to build a graph based on social network activity in a continuous manner, such that at any given time, we have a view of the social network that results from the history of the network.

Second, we use a procedure usually limited to *visualization of graphs* in the analysis of the graph. In short, we produce the layout of a graph with an extended force-directed method. We then use the layout to weight the nodes and edges of the graph a posteriori. The rationale behind this is that the information contained in a good visualization of a graph is very valuable to the human observer, and should thus be incorporated in the analysis of the graph.

We will proceed as follows. First, we introduce graph visualization (layout), and especially, force-directed layout. We will describe how the procedure operates, and how we intend to use it in our application to Enron. Then, we consider temporal aspects of the social graph. We enhance the force-directed model to account for the temporal development of the social network. Finally, we show examples of how the sampling in this new visualization space can take place.

#### 3.4.1 Force-directed Layout

A graph  $G = (V, E)$  can be visualized by drawing a circle (or other visual form) for every vertex, and a line (or arrow) for every edge connecting two vertices. We call such a visualization a *graph layout* or a *graph drawing*. It should be noted that the vertices have no a priori location. Consider a simple unweighted, undirected graph with  $V = \{v_1, v_2, v_3\}$  and  $E = \{e_{12}, e_{13}, e_{23}\}$ . When creating a layout for this graph, one can place the vertices at arbitrary locations. Figure 47 shows just three of the infinite ways to layout the given graph. One could place the circles very far apart or very close together, with equal or varying distances among each other, with straight or curved lines connecting them.

Since we do not have a priori coordinates for every vertex, the vertices must be laid out according to some criteria. Since the visualization of a graph generally aims at providing visual information about the structure of the graph,

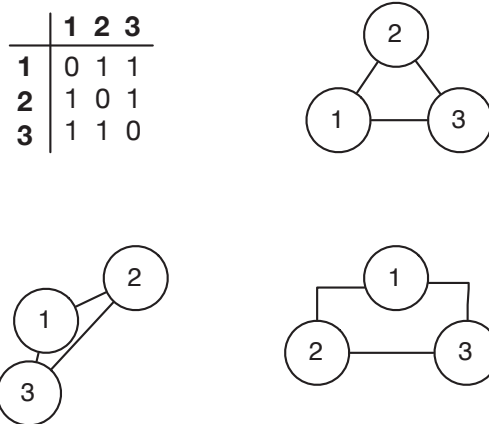


Figure 47: The graph given by the adjacency matrix (top left) can be visualized as any of the depicted graph maps. None of them is more correct than the other – they are all arbitrary (Image: F. Müller)

these criteria should be consistent with a well-readable graph layout. Battista ([138], pp. 12) considers three concepts that are relevant to the laying out of graphs: the *drawing convention*, the *aesthetic aspect*, and a certain *constraint*.

Drawing conventions are the basic rules to which the graph layout must obey. Examples are “straight-line drawing”, where edges are drawn as straight line segments, adequate for a technical network drawing, or “orthogonal drawing”, where edges are drawn as alternating horizontal and vertical segments, as in a typical organizational chart, or the concept of a “planar drawing”, where edges are not allowed to cross one another (i.e. the vertices must be laid out so that no two edges intersect). The drawing conventions can be seen as the formulation of general principles for the graph layout, depending on the type of graph and the purpose of the drawing.

Aesthetic aspects of the graph drawing are not strict rules, as in the case of the drawing conventions, but rather guidelines that are followed as much as possible in order to achieve a pleasing and well-readable layout. Examples are the minimization of intersections between edges (this is a relaxed formulation of a planar layout), the minimization of the area used for drawing the graph, which is equivalent to an effective use of available space, or the minimization of edge lengths or their variation (in an unweighted graph, we may want the edges to be of equal length, allowing some exceptions, while in a weighted graph, we may want the edge lengths to correspond to the edge weights). More general aesthetic considerations would be the aspect ratio of the layout or its symmetry.

Constraints, finally, are understood to be more specific and apply to subgraphs and their layout. Examples are the constraint that any given path should be aligned in a certain horizontal direction (e.g. flow graph), or that a subgraph

should be drawn with a predefined shape such as a star, a circle, or a square. Battista notes that many of these conventions, aesthetic aspects and constraints are computationally hard, and their exact solution may be impossible in an *efficient* layout procedure. Also, aesthetics can conflict with each other, such that one must choose one over the other depending on the application context. Any specific layout procedure is a deliberate choice in conventions, aesthetics and constraints, and its aesthetic criteria are a (deliberately chosen) subset of all aesthetic criteria.

Of the many layout procedures that exist for graphs, we will focus on *force-directed layout* procedures. In these procedures, the layout of the graph is governed by a physical model (and hence, the conventions, aesthetics and constraints are implicit in the properties of the model). The elements of the graph are physical entities, and the interaction between them is specified by physical laws that govern the magnitude and effect of forces acting on the physical entities. At the beginning of the layout procedure, the physical model is built with the components of the graph. Then, the physical laws are applied until a desired end state has been reached. The physical properties of the model in its end state are then used to create a visualization of the graph. Battista notes that force-directed procedures are often employed because (a) the physical model provides an intuitive understanding of the layout procedure, and (b) simply because the results can be very good. Many variations of force-directed procedures exist. They differ in the specifics of the model (choice of the physical entities and the laws governing their interaction) and in the algorithm used to find the desired end state.

A simple version of force-directed layout uses electrical forces and elastic springs as the core components of its model. Every vertex  $v_i$  of the graph has its equivalent in an electrically charged particle  $p_i$ . All particles are equally charged, such that they repel each other. Every edge  $e_{ij}$  in the graph has its equivalent in a spring between two particles,  $s_{ij}$ , which holds them together. In this model, the force for any vertex  $v_i$  is given by:

$$F(v) = \sum_{(u,v) \in V} f_{uv} + \sum_{(u,v) \in V \times V} g_{uv} \quad (8)$$

where  $f_{uv}$  is the effect of the spring between  $u$  and  $v$  (cohesion), and  $g_{uv}$  is the effect of the electrical force between  $u$  and  $v$  (repulsion). Assuming a two-dimensional space for the physical model (resulting from the desire to generate a two-dimensional drawing of the graph – if we wanted a three-dimensional drawing, we would also have a three-dimensional model), we can calculate the  $x$  and  $y$  components of the force  $F(v)$  as follows:

$$\sum_{(u,v) \in E} k_{uv}^{(1)} (d(p_u, p_v) - l_{uv}) \frac{x_v - x_u}{d(p_u, p_v)} + \sum_{(u,v) \in V \times V} \frac{k_{uv}^{(2)}}{(d(p_u, p_v))^2} \frac{x_v - x_u}{d(p_u, p_v)} \quad (9)$$

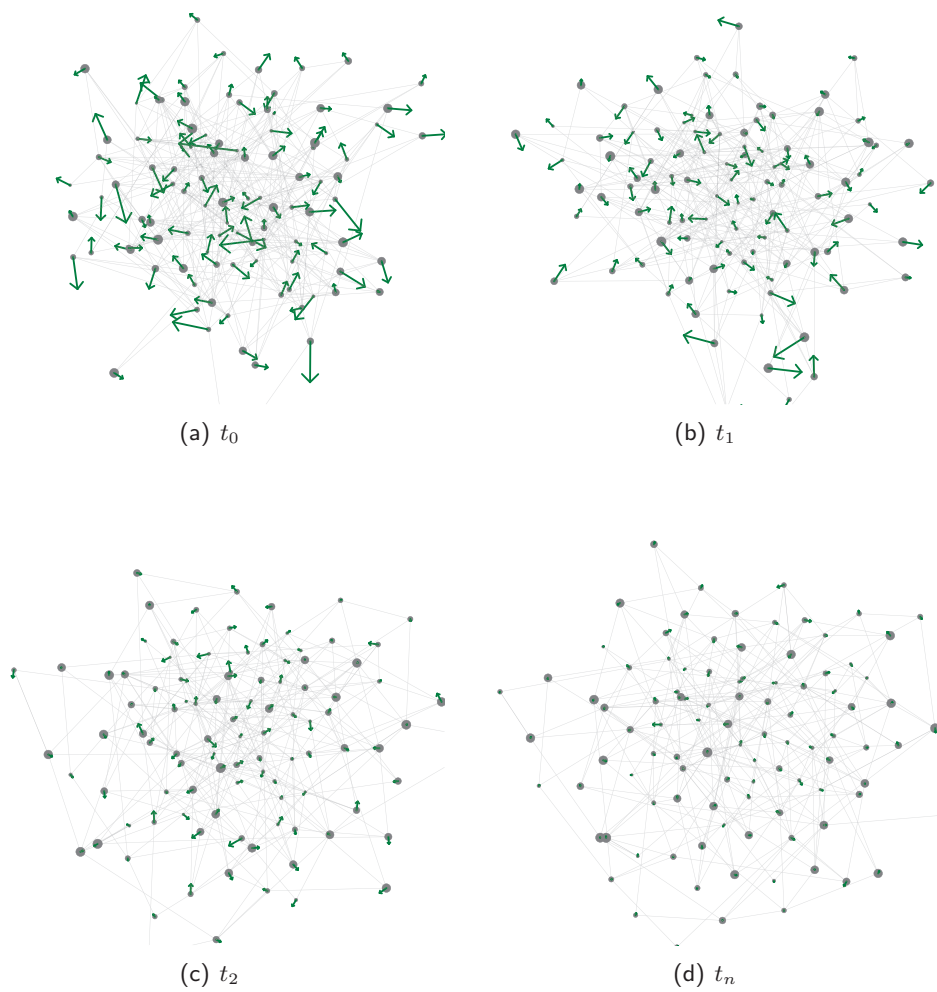
$$\sum_{(u,v) \in E} k_{uv}^{(1)} (d(p_u, p_v) - l_{uv}) \frac{y_v - y_u}{d(p_u, p_v)} + \sum_{(u,v) \in V \times V} \frac{k_{uv}^{(2)}}{(d(p_u, p_v))^2} \frac{y_v - y_u}{d(p_u, p_v)} \quad (10)$$

We can see from these equations that there are three relevant parameters to the model. The first sum stands for the  $x$  (or  $y$ ) component of the force resulting from the springs. It is determined by the *stiffness* of the spring  $k_{uv}^{(1)}$  and by the *rest length* (also called zero energy length) of the spring  $l_{uv}$ . If the distance between  $u$  and  $v$  is equal to the rest length of the spring, it exerts no force on  $v$ . The higher the stiffness, the stronger the force is effected on  $v$  when the distance between  $u$  and  $v$  is not equal to the rest length. The second sum stands for the  $x$  (or  $y$ ) component of the force resulting from the electrical repulsion. It is determined by the strength of the electrical repulsion parameter  $k_{uv}^{(2)}$  and the distance between  $u$  and  $v$ , following an inverse square law.

The combination of the repulsive force of electrical charge and the attractive force of the springs can be described as follows: the negative electrical force spaces the particles apart from one another. The springs, on the other hand, give the graph cohesion. While the rest length of the spring is the *ideal* distance between particles connected by it, its elasticity allows for a smaller or greater distance, depending on the overall connections in the model. Also, by applying varying rest lengths (e.g. depending on the weight of the edge), the meaning of the relationship between two vertices can be encoded, such as short rest lengths for strong connections and longer rest lengths for weaker connections.

The entire layout procedure operates as follows. We start with an empty model, in which we define the relevant parameters as used in Equations 9 and 10. For every vertex in the graph, a particle is created and placed at random (or using some other method) in the model. For every edge in the graph, a spring between the two corresponding particles is created. Once all vertices and edges are represented in the model, the *simulation process* is started. In every timestep  $t_n$ , the force on all particles is calculated. The impacting forces displace the particles by a certain amount. After all particles have been updated, the next step in the simulation is performed. This continues until the desired end state, the *equilibrium state*, has been reached.

The simulation is illustrated in Figure 48. In every image of the sequence, we show a drawing of the particle simulation, with the force vectors indicated as arrows. The length of the arrow corresponds to the amount of the force vector. At  $t_0$ , the particles have been placed randomly, and the initial forces and velocities have been calculated. As the simulation progresses through time  $(t_1, t_2)$ , the particles are displaced, resulting in a reduction of the forces. At  $t_n$ , the forces are minimal, and the simulation has reached its equilibrium state. Whether this state is defined as a state where the sum of all forces is zero or just below a certain threshold is a choice one can make freely. Fruchterman notes that several implementations also rely on a fixed number of simulation



*Figure 48:* Visualization of the relaxation of a particle system over time. The particles and the springs between them are drawn in gray, the velocities of the particles are depicted as green arrows whose length correlates with the amount of the velocity vector (Images: F. Müller)

timesteps, such as 100, based on the assumption that *usually*, the simulation of any such system has reached a satisfactory state of relaxation after this many steps [139]. Alternatively, one can try to express the number of steps required as a function of  $|V|$  or  $|E|$ . Instead of only considering the forces, other parameters can be included in the evaluation of the equilibrium state. As we will describe later, we have used a technique which samples the network every once in a while in order to determine whether the model is below its relaxation threshold.

The approach of the equilibrium state can more formally be described as the minimization of an energy function  $\eta$ . The criteria according to which this minimization is performed result in the specific layout of the graph. If specific aesthetic properties are required of the drawing, these can be included in the energy function. The general energy function of such a system can be described as follows:

$$\eta = \lambda_1\eta_1 + \lambda_2\eta_2 + \dots + \lambda_k\eta_k \quad (11)$$

where  $\eta_i$  is a specific energy function and  $\lambda_i$  is a weighting factor. Coming back to the illustration in Figure 48, where we have indicated how the overall force acting on the particles is reduced over time, we have a simple energy function with  $\eta$  being the overall force on every particle and  $\lambda$  equal to 1. We could also choose to treat the electrical and spring force component separately (with different  $\lambda$ ), and even introduce further minimization criteria, such as the number of edge crossings, if we wanted to minimize the latter.

Force-directed methods for graph drawing produce good results, even if they introduce an element of arbitrariness. While it is very unlikely that two simulation runs of the same graph produce the same layout, the results are likely to be *equally good*. We have only introduced a simple model, and it should be noted that several more complex models exist. Depending on the application, they are better suited for obtaining visually pleasing results [139] [140] [141].

### 3.4.2 Temporal Aspects of Social Graphs

Suppose that for a small group of individuals, we have data about who communicated with whom (and when) during an entire year. If we build a social graph, in which every individual is represented by a node, and every communication of the entire year is represented by an edge (or a fraction of an edge), we construct an accumulated (or summarized) view.

Consider that individuals A and B have exchanged a little more than one message every day in January, and that C and D have exchanged messages on a weekly basis over the entire year. Suppose further that in the drawing of the graph, we take the number of communications between two nodes to be the weight of the edge, and draw it with an according thickness. The line between A and B will be similarly thick as the line between C and D, although the manner in which A and B communicated is very different from the manner in which C



and D communicated.

Of course, one could construct a separate social graph with the data for every separate week, or every separate month. In the graph for January, A and B would be connected by a thick line, while in the graph for any other month, they would not be connected at all. While this is better than in the case of the aggregate view, it would still lack an important aspect, namely, that communication between people echos in the future. The aggregate view of social data makes it disconnected from its past – it seems intuitive that if I A has had a lot of communications with B in January, but not a single one in February, then A and B are still somewhat related in February (the absence of communication, compared to the previous month, may be relevant). And if in April, A and B take up communication again, it's not the same as if A started to communicate with a new individual E – *A and B have history*, and it is likely that this history is relevant for their current communications.

It is for this reason that in our analysis of the Enron social network, we want to consider the evolution over time. When we look at the drawing of a social graph, it always represents a snapshot at a specific time. However, we believe that this snapshot should – in its generation – take into account the state of the social network at earlier times. This principle, which we call the *reaction-diffusion* model (RD model), is illustrated by its application to the example of the year-long communication between A, B, C, and D mentioned earlier in Figure 49.

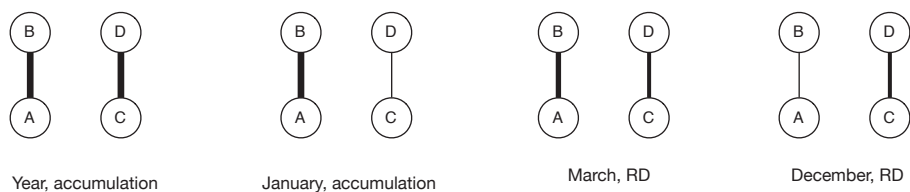


Figure 49: Drawings of the social graph for a communication network using aggregation (top) and a reaction-diffusion model (bottom) (Image: F. Müller)

In general, a reaction-diffusion system is a system in which the concentration of certain substances are governed by an environment-dependent reaction process which increases the concentration, and a time-dependent diffusion process which decreases the concentration. If a substance present with a certain concentration does not find an environment for reaction, its concentration will decrease continuously, until it has vanished. We apply a similar model in the construction of the graph of a social network. As agents appear and communications occur (reaction), the graph is built with its respective components, vertices and edges. As we progress through time, the graph components are subject to diffusion, i.e. their importance will decrease over time, and should they not be part of further interactions, they will eventually disappear. The

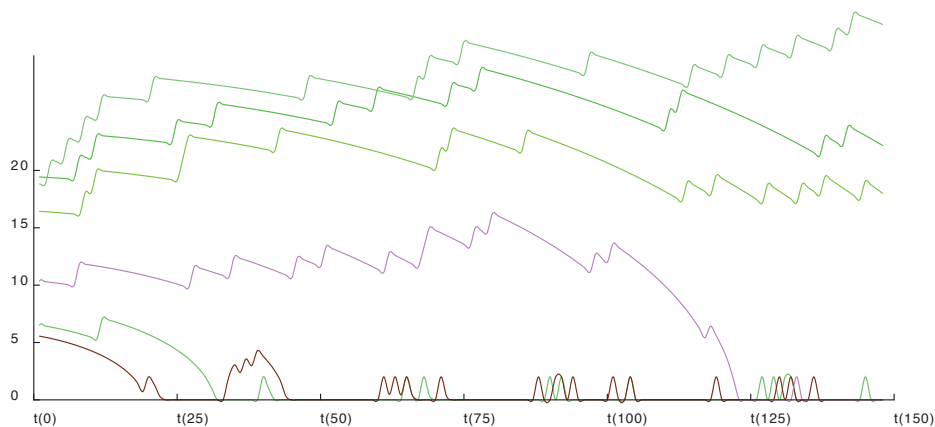
reaction and the diffusion, in our model, do not relate to a concentration of substances, but to the mass of the particles by which the communication participants are represented (see also Section 3.4.1). The diffusion for the mass and the strength are calculated as follows:

$$\Delta m = \frac{k_m}{m} \cdot \left(1 + \frac{a}{a_{max}}\right) \quad (12)$$

$$\Delta s = \frac{k_s}{s} \cdot \left(1 + \frac{a}{a_{max}}\right) \quad (13)$$

where  $n$  is the number of agents involved in the communication,  $k_m$  is an empirical parameter for mass reduction, and  $k_s$  is an empirical parameter for strength reduction. They determine the amount by which the mass or strength are reduced in every timestep. Figure 50 shows the development of the mass for a particle simulation with 6 particles using the reaction-diffusion model. The particles have an initial mass of between 5 and 20. As the simulation progresses, the particles continuously lose some of their mass in every time step. At certain times, discrete growth occurs, by which the mass of the particle(s) increases considerably. As evident from the figure, heavier particles lose mass less quickly. This aspect can be steered by the constant representing the maximal age.

The aim of using the reaction-diffusion model is to account for the relevance of temporality in social relationships. The effect of agent communication (in the form of mass increase) is realized at the time the communication takes place, and over time, this effect fades away. If an agent ceases to communicate, it will eventually be forgotten, no matter how prominent its role once was.



*Figure 50:* Illustration of the mass development of various particles: increases occur when communication between two vertices takes place, the constant decrease in mass is due to the reaction-diffusion model (Image: F. Müller)

### 3.4.3 Sampling the Distance in Visualization Space

It was stated in the introduction to this section that we are interested in the formation of clusters within the social network. Clustering algorithms have briefly been introduced in section 3.2.1. We have seen that they are all based on some distance metric, which in the case of a graph is usually a measure such as closeness or betweenness. Instead, we choose to take the visualization of the graph as basis for the distance measurement. The rationale behind this is as follows. The force-directed layout of the graph can be said to *encode* ([138], p. 304) the aesthetic criteria of the graph layout. Our version of the force-directed layout incorporates the temporal evolution of the social network and introduces a reaction-diffusion model to account for the diminishment of importance of actors that stop communicating. It has been said that the aesthetic criteria follow the goal of a well-readable graph, and this good readability can be interpreted to 'reveal information'. A good layout is not only visually pleasing, it also expresses meaningful features of the underlying graph (i.e. the underlying social network, in our case) in the visual domain.

More specifically, our approach is to use the particle simulation as a basis for deriving a weighted version of the Enron graph. The masses of the particles are used to weight the vertices, and the distances between the particles are used to weight the edges. We then perform an agglomerative clustering on the graph. We expect that the organizational structure evident in the visualization of the Enron graph will find its equivalent in the clustering performed on the visualization-weighted graph.

## 3.5 System Architecture and Implementation

In this section, we describe the architecture and the implementation details of our approach, which was described in the previous section. We start by the describing the overall system architecture. We then detail the assumptions we have made in order to run the particle simulation on the Enron corpus, and how the results of the simulation are applied to the Enron social graph. Finally, we provide the results which we have obtained, in the form of visualizations of the Enron social network and of an automatically generated organizational chart, which is compared to a manually researched organizational chart.

### 3.5.1 System Architecture

The system consists of four components, responsible for data extraction, simulation, clustering, and visualization. The data extraction component retrieves the corpus data from the database and exposes it in the form of a Java object. A collection of these objects is the input for the simulation component, which contains the physical force relaxation simulation and the reaction-diffusion model. The simulation component outputs simulation snapshots at a user-defined frequency (e.g. weekly snapshots). These snapshots contain the state of the Enron social graph at the given time as resulting from the simulation model. More specifically, it contains a map of the social graph, where every vertex has a location in two-dimensional space, and an adjacency matrix that represents the strength of the springs between the connected particles. The simulation snapshots are used by the two remaining components, the clustering and the visualization component. The clustering component implements the weighting of the Enron social graph with the Euclidean distances measured from the two-dimensional representation in the snapshot. It then clusters the weighted graph to derive an organizational chart of the entire Enron network. The visualization component operates on a set of simulation snapshots and allows the navigation through time. It has been implemented in a week-wise manner. For any given week, the graph can be weighted with several measures, and is displayed accordingly. In addition, the clustering algorithm can be applied interactively. Within a certain parametrization (weighting, clustering), an animation of the development over the weeks can be displayed. Figure 51 illustrates the four components and their interfaces. In the remainder of this section, the four components are described in detail.

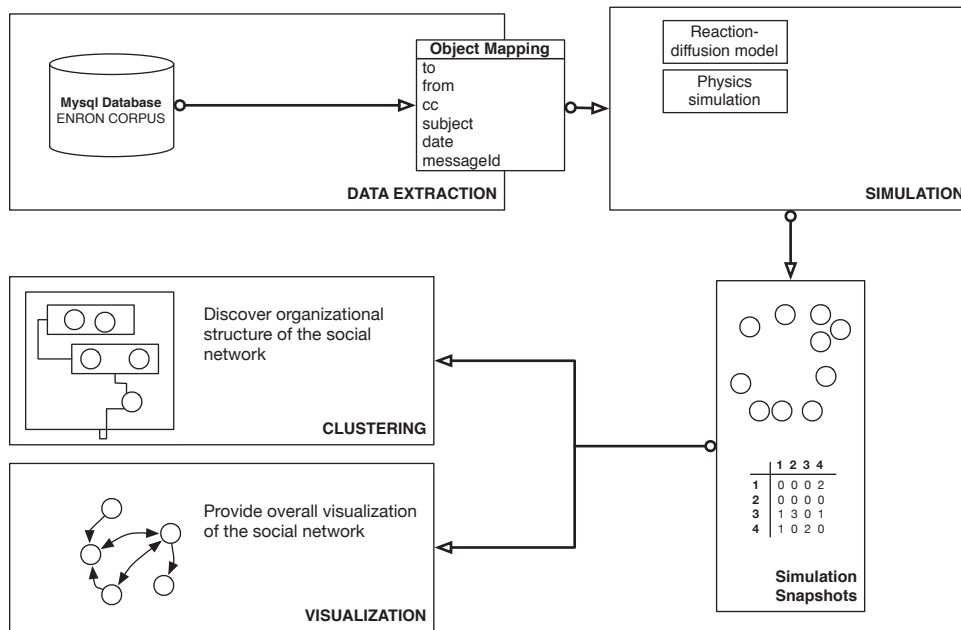


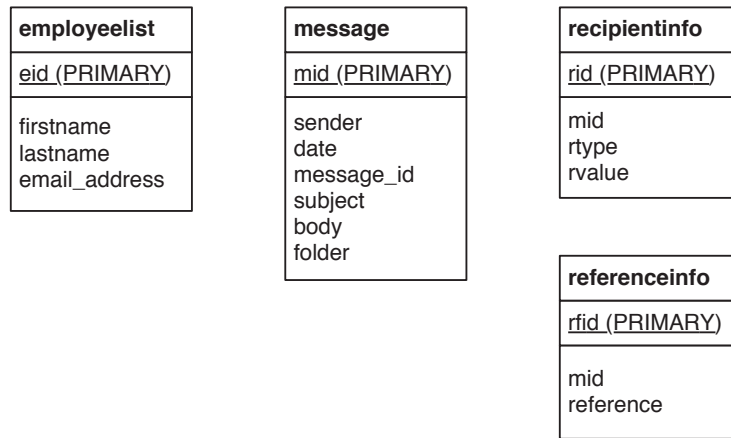
Figure 51: Overall System Architecture (Image: F. Müller)

### 3.5.2 Database and Data Extraction

Our work is based on the dataset provided by Shetty and Adibi [126]. Starting from the dataset provided by Cohen [123], they have removed duplicate emails and provide a relational database containing 252,759 messages, and in addition, a list of all the custodians. Their database is organized in four tables: one with the list of the custodians, one with a list of sent messages, one with a list of receivers (referencing the sent messages via a unique id), and one with reference information. The table with reference information contains a subset of all available messages, namely those that are a reply to or a forwarding of another message.

The table `employeeelist` contains a list of 151 custodians, and one single email address for every custodian. Several custodians, however, have several email addresses. The `email_address` entry for Vince J. Kaminski in the `employeeelist` is `j.kaminski@enron.com`. However, when searching the database for sent messages originating from an address that contains the string 'kaminski', we obtain the results displayed in table 9. This is also documented by Zhou et al. [128].

As can be seen, the associated address in the `employeeelist` table only accounts for one fifth of the emails sent by Vince Kaminski. We have queried the database for addresses that could be associated with a custodian, and have found that in many cases, custodians have several email addresses. We have



```

SELECT message.sender,recipientinfo.rvalue FROM message
LEFT JOIN recipientinfo on message.mid=recipientinfo.mid
  
```

Figure 52: Structure of the MySQL database containing the Enron Emails as provided by Shetty and Adibi. The SQL statement shows how sender as well as recipient information of a message can be obtained (Image: F. Müller)

Address	Scheme	Number of sent messages
vince.kaminski@enron.com	FF.LL@	4366
j.kaminski@enron.com	M.LL@	1219
kaminski@enron.com	LL@	253
vkaminski@aol.com	n.a. @	163
j.kaminski@enron.com	M..LL@	4
vince.j.kaminski@enron.com	FF.M.LL@	1
vkaminski@palm.net	n.a.	1

Table 9: Vince J. Kaminski's email accounts. The *Scheme* column lists the derived naming scheme found in all multiple addresses for custodians, where FF stands for the spelled out first name, M for the initial of the middle name, and LL for the spelled out last name.

derived a naming scheme according to which Enron seems to have assigned email addresses, and the listing for Kaminski in table 9 exemplifies all of them. We have manually checked a total of 1,000 candidate addresses, and have assigned multiple addresses to individual custodians if they conform to the derived naming scheme. Note that this applies only to email addresses *@enron.com*, email addresses from other domains were not included. While in some cases, they can be related to a custodian with rather great certainty (as in *vkaminski@aol.com*), it would have been more speculative in other cases, which is why we have not considered them all. We have increased the ratio of addresses per custodian from 1.0 in the original database to 1.4. The assignment of multiple addresses is used in the simulation to model the various addresses as one single entity.

The data extraction module is responsible for querying the database and providing the email data to the simulation module. Every email is encapsulated by a `EnronEmail` object, which contains all the header information of the email. The relationship between the database and the email object is illustrated in Figure 53.

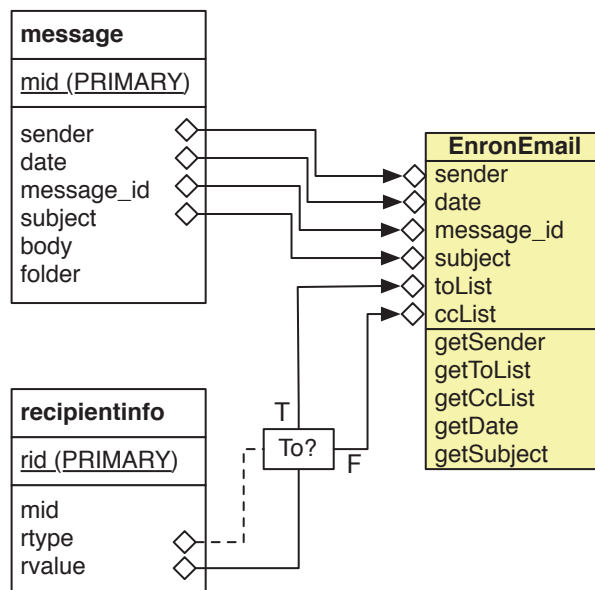


Figure 53: Database to business object mapping (Image: F. Müller)

### 3.5.3 Simulation Model

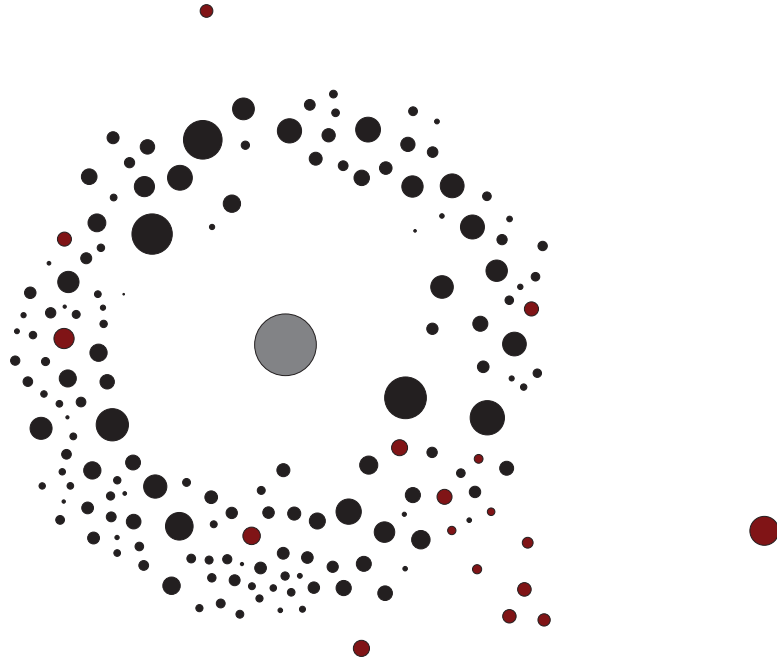
As has been stated in section 3.4.1, we use a physical model with electrical forces and springs to layout the Enron email corpus. The relationship between the emails and the model are as follows: all the email addresses of a custodian (which may be one or more) correspond to one and the same particle, and every email between any number of senders and recipients corresponds to a spring that is created between the respective particles. Within the entire database, there are 17,568 unique addresses from which emails have been sent, and 68,214 unique addresses that have received emails. If we limit the addresses to the domain enron.com, we have 6,066 unique senders and 29,049 unique recipients. Experiments have shown that representing each unique address by a particle is not feasible for two reasons. First, the computational complexity of simulating tens of thousands of particles is prohibitive. Second, the huge number of vertices in the resulting graph makes it very hard to draw such that information is evident. We have therefore come up with a simplified representation for the simulation.

1. All addresses belonging to a custodian are mapped to one and the same particle representing that custodian
2. All addresses in the domain enron.com that are not assigned to a custodian are mapped to one and the same particle, representing all of Enron corporation
3. All addresses outside of the domain enron.com which are on a list of the 15 most frequent domains are assigned to one particle representing each domain
4. All addresses outside of the domain enron.com which are not on the list of the most 15 frequent domains are ignored
5. The particle representing all of Enron corporation is assigned a fixed position at the origin of the two-dimensional coordinate system and is immobile

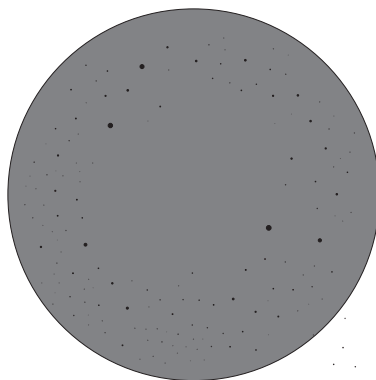
A visualization of the particle simulation emphasizing this structure is provided in figure 54. The light gray particle in the center of the drawing represents all of Enron corporation minus the custodians. The black particles around it represent the custodians. The red particles represent the most frequent outside domains. The size of the particles is proportional to their mass. The mass, in turn, represents the frequency of communications (subject to the RD model). The Enron particle is displayed in its own scaling, since it has a very high mass (it represents thousands of employees). A drawing proportional to its mass is shown in Figure 55.

The simulation is based on a physics library provided by Bernstein [142]. It provides a particle system, in which particles and forces can be created. Every





*Figure 54:* Enron simulation. In the center is a big particle representing all Enron employees not considered for special attention (medium gray). The black particles are the 160 key employees under special attention. The red particles are external domains (e.g. dukeenergy.com). The size of the central particle is disproportionate to the size of the other particles, see also Figure below (Image: F. Müller)



*Figure 55:* As in the Figure above, but the size of each particle is directly proportional to its mass (Image: F. Müller)

particle has a position, a mass and an age. The forces are introduced either as *attractors* or as *springs*. Attractors operate between a pair of particles can be either positive (attractive force) or negative (repulsive force). Their strength follows an inverse square law. Springs are defined between two particles, and have a zero energy length, a strength (spring constant) and a damping factor. Equation 14 shows the force effective on any particle resulting from the mutual pairwise repulsion of the particles.

$$\vec{F}_{rep}(p_i) = \sum_{i \neq j} \frac{c \cdot m_i \cdot m_j}{|\vec{r}_i - \vec{r}_j|^2} \quad (14)$$

where  $c$  is a constant defining the strength of the repulsion,  $m_i$  and  $m_j$  are the masses of the particles, and  $\vec{r}_i$  and  $\vec{r}_j$  are the position vectors of the particles. Equation 15 shows the force effective on any particle resulting from the force of the spring between it and any other particle.

$$F_{att}(p_i) = \sum_{i \neq j} -(|\vec{r}_i - \vec{r}_j| - d_{ij}) \cdot k + (\vec{r}_i - \vec{r}_j) \cdot (\vec{v}_i - \vec{v}_j) \cdot c_{damp} \quad (15)$$

where  $d_{ij}$  is the rest length of the spring between the two particles,  $k$  is the spring constant,  $\vec{v}_i$  and  $\vec{v}_j$ , are the velocity vectors of the two particles, and  $c_{damp}$  is the damping constant of the spring. The first term of the equation describes a classical spring model, the second term describes the velocity-dependent damping force of the spring.

The simulation module operates as follows. It is given the *time window* which it should cover, indicated by the first and the last week of the period (e.g. 1999-41 to 2001-22). Starting with an empty particle system, it iterates through the weeks of the time period. In every week, it first applies the reaction-diffusion model to account for the diminished importance of the previous communication. For every particle and every spring, the reaction-diffusion ('ageing') function as described in section 3.4.2 is applied, and the weight is decreased accordingly. Second, it processes all the emails of the current week<sup>52</sup>, updating the weights and connections of the particles. If for a given address, the corresponding particle already exists, its weight is increased. If the corresponding particle does not exist, it is newly created. The weight increase or the initial weight (*base weight*) of a new particle depends on the message for which it is updated or created. For the sender, the value is 1, and for the recipient(s), it is  $\frac{1}{n_{to}}$  for direct recipients, where  $n_{to}$  is the number of direct recipients, and  $\frac{1}{n_{cc}}$  for carbon copy recipients, where  $n_{cc}$  is the number of carbon copy recipients.

If a new particle is created, there are two possibilities for placing it in the model. If another particle occurring in the respective email is already present

---

<sup>52</sup>We have found situations where emails with a very high number of recipients (several hundreds) lead to a catastrophic failure of the particle simulation upon their introduction. We have thus filtered out messages with more than 100 recipients.

in the simulation (the check starts with the sender, continues with the direct recipients, and finally the copy recipients), the new particle is placed randomly in the *neighbor field* (the close neighborhood of the particle as defined by the parameters, see Table 10) of the other particle. If none of the addresses in a communication is already present as a particle, then the sender particle is first placed randomly in the particle field, which is defined as an area of a certain size in which the particles reside initially (see Table 10). The other particles are then placed randomly in its neighbor field. Whenever a new particle is introduced, a negative attractor (corresponding to the electric repulsion) is created between it and every other particle in the system. For every communication link between two particles ( $a$  is the sender and  $b$  is the recipient, or v.v.), a spring between these two particles is either created (if no such spring had existed), or the existing spring between the particles is updated. The springs all have a rest length of 300 units and the default strength ( $k_s$ ). When a spring is updated, its strength is changed by an increment equivalent to the default strength.

Once all the particles of the week have been processed, the simulation is run until it has reached its equilibrium state. The equilibrium state is defined as a state in which the average velocity  $\bar{v}$  of all the particles in the system is below a threshold  $\epsilon = 0.1$ . As soon as the equilibrium state is reached, the configuration of the particle simulation is persisted for future use (this is the *snapshot*). This configuration includes the names and masses of all particles and their positions, as well as an adjacency matrix describing the strength of the springs between the particles. The procedure of the application of the RD model and the processing of the communication is illustrated for four messages in Figure 56.

The simulation could be run on any number of weeks. We have chosen to create a reference simulation spanning from the first week of 1999 to the week 42 of 2002, a total of 188 weeks. This results in 188 snapshots of the simulation, each detailing the particles, positions, and masses. These snapshots are used in the clustering algorithm described in section 3.5.4. They also provide the basis for the interactive visualization tool described in section 3.5.5.

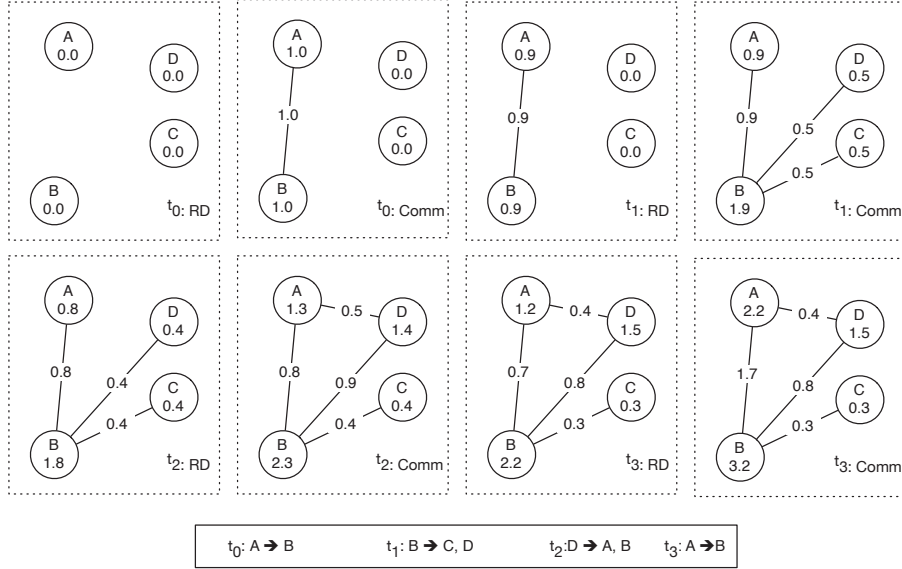


Figure 56: Progress of the particle simulation over four time steps. For illustrative purposes, we have chosen  $\Delta m$  and  $\Delta s$  of the RD model to be constant at 0.1. For every timestep  $t_i$ , the RD and the Communication processing phase are shown (Image: F. Müller)

Parameter	Value
Particle field	$[-1000, 1000]$
Neighbor field	$[-100, 100]$
Particle system drag	0.1
Particle Base weight ( $w_b$ )	1
Sender weight	$1 \times w_b$
Direct recipient weight	$w_b/n_{to}$
Copy recipient weight	$w_b/n_{cc}$
Attractor force	-100
Default spring strength ( $k_s$ )	0.02
Initial spring rest length ( $d_s$ )	300
Spring strength update	$k_s$
Spring damping ( $k_{sd}$ )	0.02

Table 10: List of simulation parameters

### 3.5.4 Weighting and Clustering

As introduced in section 3.4.3, we use the result of the simulation – the drawing of the graph – to create a weighting for the Enron social graph. Every vertex  $v_i$  of the social graph has a corresponding particle  $p_i$  in the physical model, and every edge  $e_{ij}$  of the graph has a corresponding spring  $s_{ij}$ . The weighting of the vertices and edges corresponds to the masses of the particles and the distances between them:

$$w(v_i) = \text{mass}(p_i) \quad (16)$$

$$w(e_{ij}) = \sqrt{(s_{ix} - s_{jx})^2 + (s_{iy} - s_{jy})^2} \quad (17)$$

where  $w(v_i)$  is the weight of the vertex and equal to the mass of the particle, and  $w(e_{ij})$  is the weight of the edge and equal to the Euclidian distance between the two particles. Note that the graph we obtain has density  $E = 1.0$ , since an edge is introduced between every pair of vertices. Also note that the particle representing all other Enron employees is not transferred to the graph. We then apply an agglomerative clustering algorithm to the graph as follows.

The vertices of the graph represent the custodians and the external domains. At the beginning of the clustering, every vertex is an individual cluster. We initialize a distance threshold  $d$  to zero. Then, we steadily increment  $d$ . For every value of  $d$ , we check whether two vertices  $v_i, v_j$  exist such that  $w(e_{ij}) \leq d$ . Since every vertex is connected to every other vertex, we introduce an additional constraint, consisting in the condition that a spring must exist in the simulation between  $v_i$  and  $v_j$  (i.e. they have communicated, and their communication is still notable). If both conditions are satisfied, the two vertices are agglomerated into a cluster. With an increasing value of  $d$ , more and more vertices are clustered together. Through the introduction of the communication constraint, the clustering will not necessarily result in one single cluster spanning all vertices of the graph. Therefore, an upper limit for the distance threshold is defined. As a value, we have taken the maximum distance between any two particles, i.e. the highest weight of any edge in the graph.

The resulting agglomerative clustering is illustrated in Figure 57. The  $x$  axis corresponds to the vertices of the graph, the  $y$  axis corresponds to the distance threshold  $d$ . While at  $d = 0$ , all vertices are located in their individual cluster, these are agglomerated as  $d$  increases. As can be seen, the great majority of the vertices is agglomerated into one single cluster eventually. In the visualization, the clusters are sorted such that all vertices eventually being agglomerated in the largest cluster are to the left, while 'disconnected' vertices – vertices that remain in their individual cluster, since they have had no relevant communication with other vertices – are all located on the right side. The largest value of the  $y$  axis corresponds to the largest distance between any two vertices in the graph.

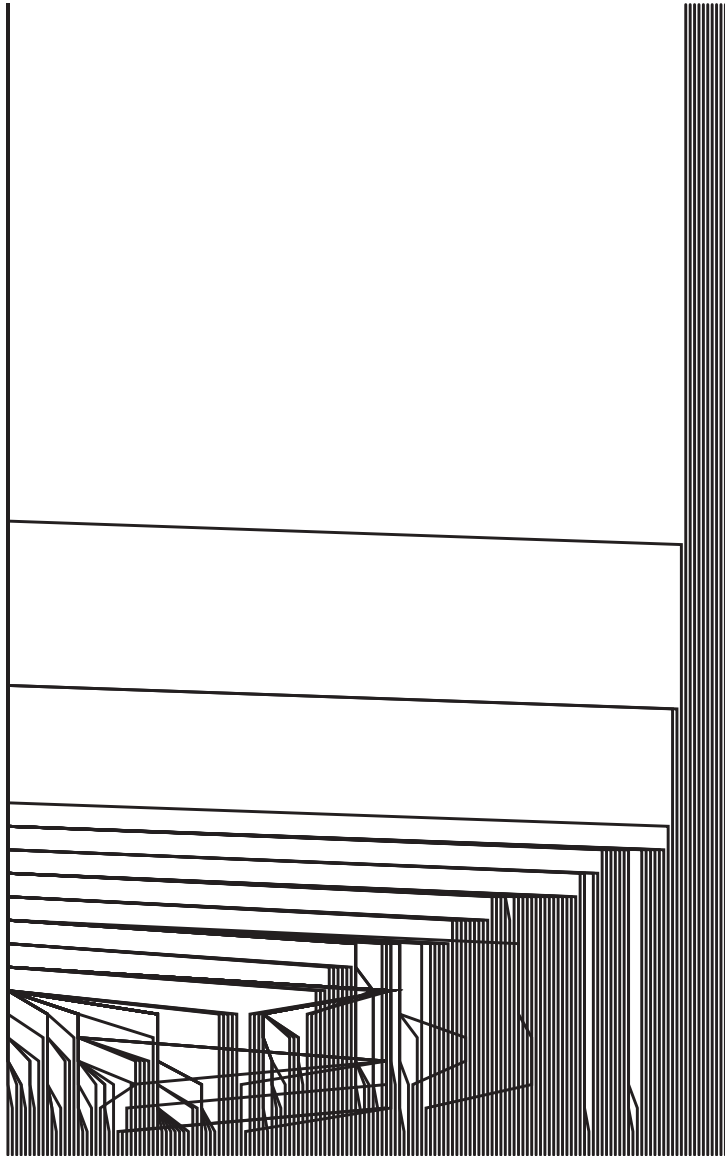


Figure 57: Agglomerative clustering of the Enron graph for week 2001-34 (Image: F. Müller)

While Figure 57 provides an illustration of the clustering of the Enron social graph, Figure 58 shows a tool that was developed to inspect the clustering. In comparison to the illustration, the view is rotated 90 degrees, such that the  $x$  axis corresponds to the distance threshold and the  $y$  axis corresponds to the individual vertices. Every cluster is assigned a random color, and a colored block represents the cluster membership of any vertex for a given distance threshold.

The formation of the clusters can be observed, and the full member list of any cluster can be displayed by selecting an individual cluster through a click on the respective color.

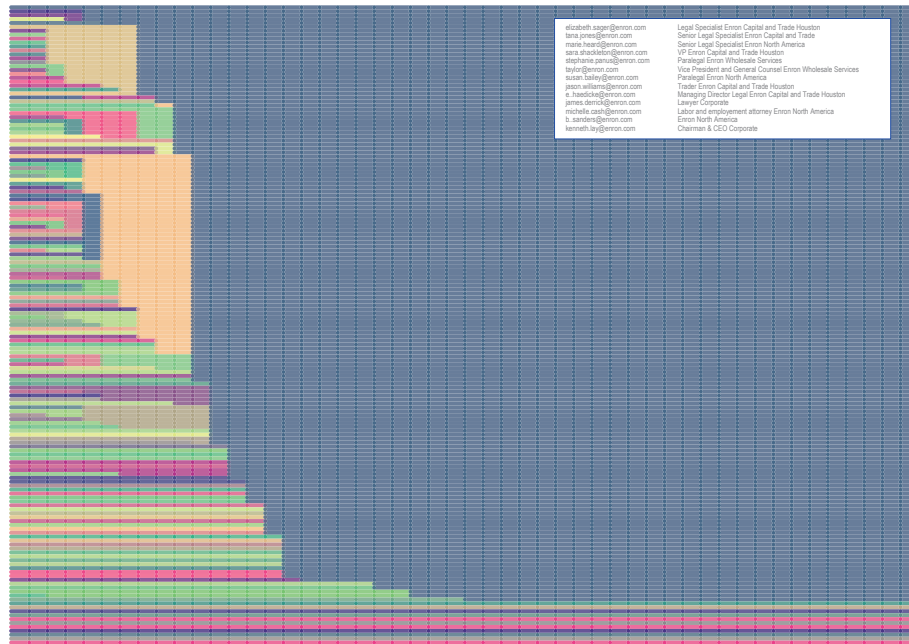


Figure 58: Interactive tool to view the clustering results. The  $x$  axis corresponds to the distance threshold, the  $y$  axis to the custodians (Image: F. Müller)

The goal of the clustering is a partial derivation of the organizational structure of the underlying social network. In order to evaluate our results, we have tried to reconstruct as much as possible of the organizational structure of Enron. As sources, we have used the Enron Ex Employee Status Report by Shetty and Adibi [143], the Employee Committee's complaint against J.D. Arnold (U.S. Bankruptcy Court Case No. 01-16034-AJG [144]), the report by Creamer et al. [135] and a full-text search of the database (USC corpus), where many emails contain indications of organizational adherence, e.g. in signatures and address book entries. A table listing all the custodians and their associated organizational status is included in Appendix A.3. Apart from the rank or function of individual employees (such as *Vice President* or *Trader*), we were interested in the broader organizational attributes.

We have identified the following organizational units: Capital and Trace (ECT), Corporate, Transportation Services (ETS), Energy Services (EES), Enron North America (ENA), Enron Online, Enron Wholesale Services, and Enron Global Markets. It should be noted that ENA is actually the successor unit to ECT, of which a part was renamed to ENA after a reorganization. How-

ever, we have not detailed information about the details of this reorganization. Also, the label ECT was still being used throughout the entire time for which there are email in the corpus. Therefore, we have maintained the separation in our analysis. The close relationship is respected in that both units are colored similarly.

Each of these units has been assigned to several employees. For the ENA unit, more about detailed substructure is known: it contains ENA Real Time, ENA Northwest, ENA Middle Market, ENA Volume Management, ENA Fundamental Analysis, ENA California, ENA Southwest and ENA Services. The source of this sub-unit structure is a report by McCullough Research [145]. Unfortunately, only some custodians that have been identified as belonging to ENA have also been associated with a sub-unit in the report.

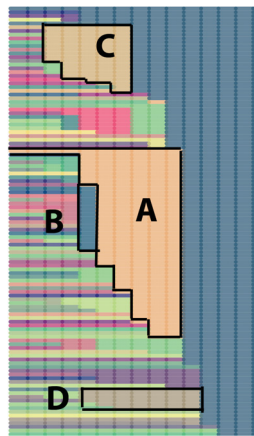


Figure 59: Use of the clustering tool to identify clusters and subclusters (Image: F. Müller)

Based on the clustering, we have manually created several partial organizational charts of the custodians. We have used the clustering interactive tool for this purpose. Figure 59 shows a screenshot of this interface. Clusters and their subclusters can be easily identified. In the example, four clusters are labeled: A and its subcluster B, as well as C and D. The derived organizational charts include all clusters with at least three members.

Two are discussed here, the one derived from the snapshot of week 52 of the year 2000, and the one derived from the snapshot of week 52 of the year 2001. They are shown in Figure 60 and Figure 61, respectively. The organizational units are color-coded. For some employees, no assignments could be made, in which case they are not colored. The same applies to external domains, which have not been assigned a unit.

In both charts, a group with members predominantly from ETS can be identified (top left). They are clustered together with several external domains, which are mostly gas and electricity companies. ETS was responsible for the



operation of the pipeline and electricity network, which explains why these external domains (using Enron's infrastructure) communicated with ETS. In the 2000 chart, Dasovich and Shapiro form a subgroup in the ETS cluster, together with the external domain govadv.com. The two employees were responsible for regulatory affairs and government relations. It seems that the operation of the physical infrastructure required frequent coordination with them.

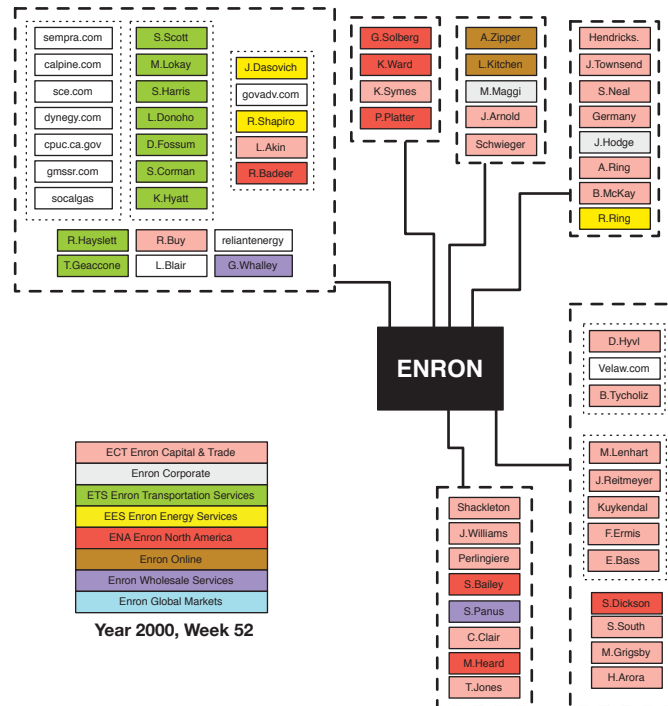


Figure 60: Partial organizational chart of Enron derived from the simulation procedure, week 52 of year 2000 (Image: F. Müller)

In the 2001 chart, we can identify a group consisting only of ENA employees (middle right). We have found a specific sub-unit for all of them: Semperger, Scholtes and Crandall were assigned to ENA Northwest, Salisbury to ENA Real Time, Platter to ENA California, and Gang and Motley to ENA Southwest. A comparable, less extensive concentration can be found for a 4-constituents group in the 2000 chart (top, second from left), where Solberg was assigned to ENA Real Time, Ward to ENA Middle Market, and Platter to ENA California.

In both charts, a group in the bottom seems to have a legal focus. In 2000, the group consisting of Shackleton, Williams et al. has 8 members. Six of them are in the role of either legal specialists (Perlingiere, Clair, Heard, Jones) or paralegal (Bailey, Panus). In 2001, the group of Shackleton, Panus et al. (containing 2 subgroups) has a total of 12 members. Nine of them have a role with a legal focus, among them Headicke as 'Director Legal' and Taylor as

'General Counsel'.

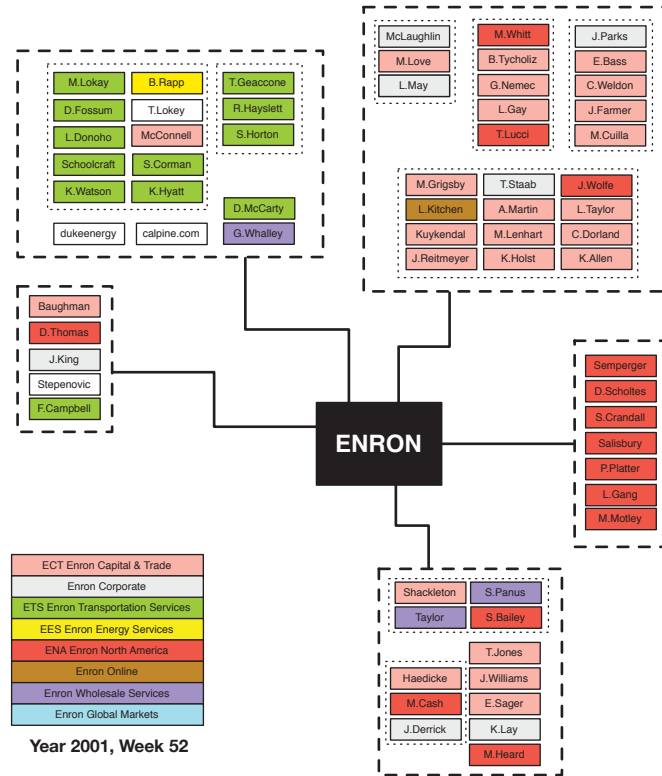


Figure 61: Partial organizational chart of Enron derived from the simulation procedure, week 52 of year 2001 (Image: F. Müller)

### 3.5.5 Interactive Visualization

In the context of electronic memory, visualization has received considerable attention. This includes specific work for novel email interfaces. Most current email interfaces are relatively unsuitable for providing e-memory interfaces to personal communication. They are usually list-based, i.e. the messages are displayed in a chronological (or otherwise sorted) list, and details about the currently selected message are displayed in a separate view component. While search functionalities enable users to retrieve messages that match specific criteria, we have a lack of associations between messages.

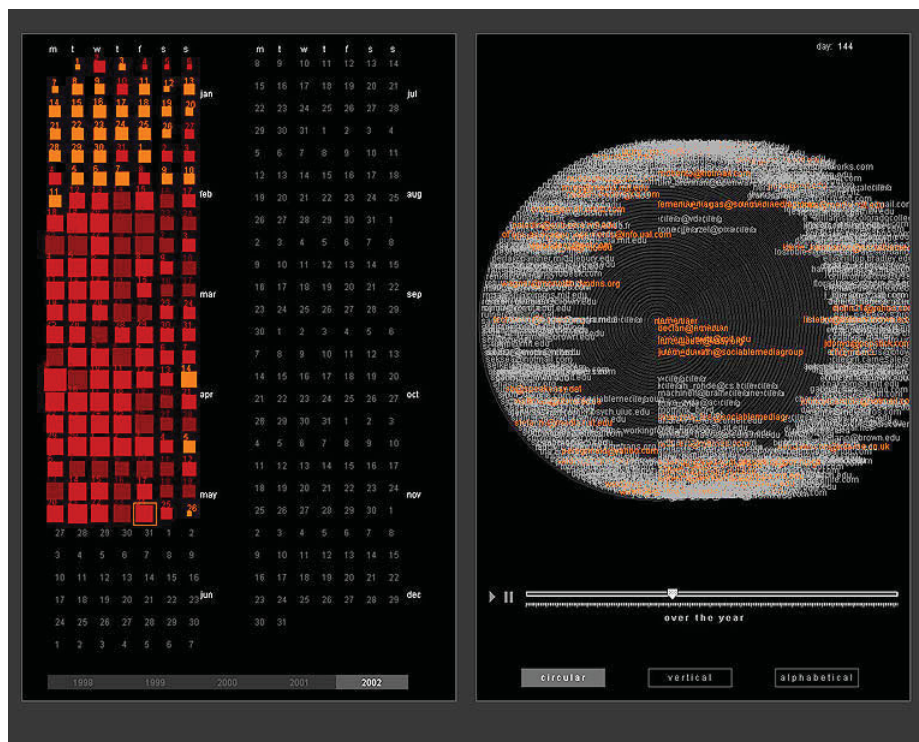


Figure 62: PostHistory Email Interface. Left: calendar interface for time period selection. Right: social network view of selected communications and playback interface (Image: Fernanda Viégas [146])

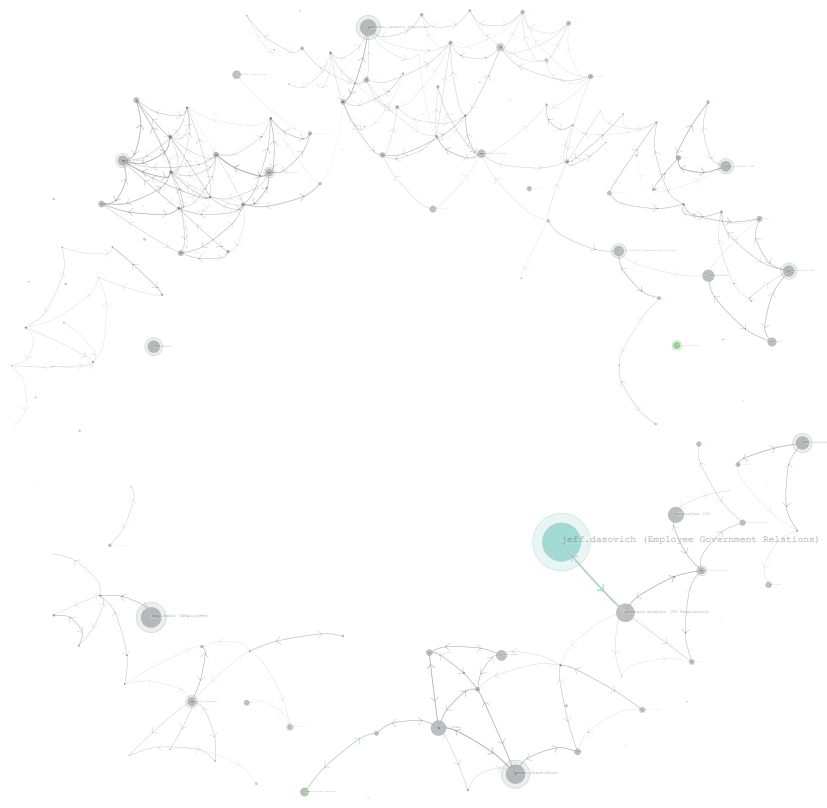
Several new approaches have been suggested, and they mostly provide a view that is organized in part or in full around the individuals that participate in the communication – it is basically a view of the social network implicit in the email network. Frau et al. [147] have proposed a dynamic email interface 'Mailview' which displays plots of email communication over time. It focuses on visually aggregating existing message attributes such as size or folder. Viégas has developed several visualizations for email, including the PostHistory interface [148] [149]. A visualization of the system is depicted in Figure 62. It presents

users with an overview of their email communications through the use of a timeline overview in the form of a calendar and a visualization of the contact network of the user. The organization around a time stream, quite intuitive by itself, has prominently been suggested by Freeman and Gelernter in their 'Lifestreams' project [150]. Depending on time window and contact selections, the communication with one or several persons is displayed over time in the calendar view, and relevant social network context is displayed in the contact view. Heer [151], while working on the Enron corpus, has developed an interactive social network viewer for the Enron graph. It allows the interaction with a real-time visualization of the graph, individual messages can be selected via their participants.

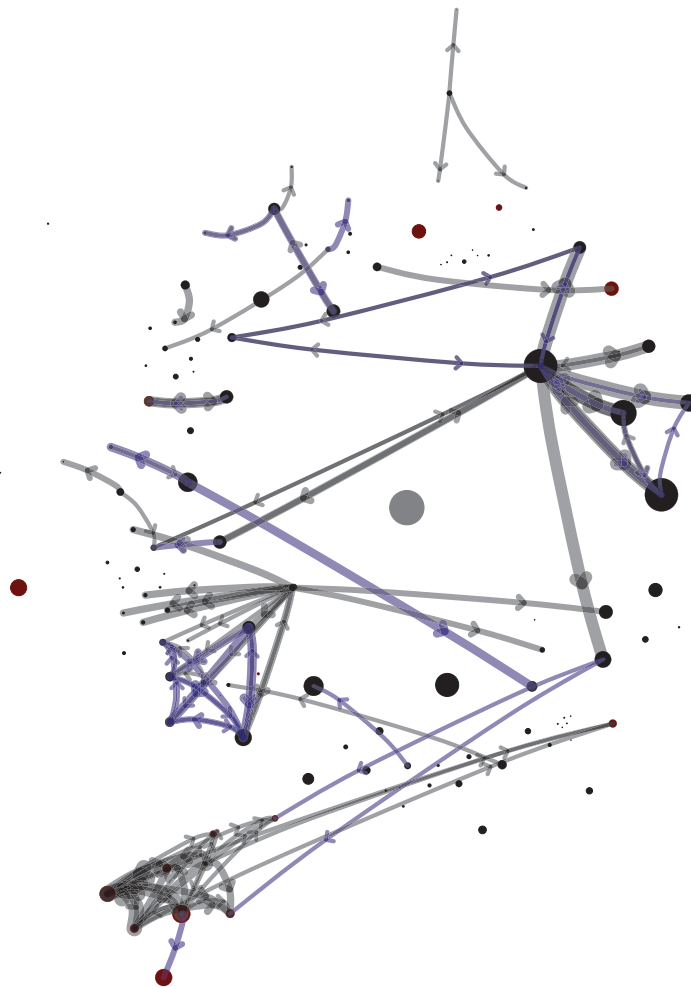
Based on our simulation model, we have developed an interactive visualization tool. The focus does not lie on individual messages, i.e. it is not an email viewer. Rather, it is an interactive social network explorer. It uses the snapshots of the particle simulation as its data basis. The user can select which time period to view (in the case of our Enron implementation, any week included in the simulation run) and chose between two display modes: either, the user can view the communication that occurred throughout the entire graph, or only the communication that has (implicitly) led to the formation of the clusters. Figure 63 shows the Enron social graph, together with the communication that was considered at a certain distance threshold. The vertex corresponding to J. Dasovich has been manually selected and is thus highlighted.

Figure 64 show the Enron social graph for the week 32 in the year 2000. All communication activity is shown, direct emails are indicated in gray, copy emails are indicated in blue. The weight of the arrows corresponds to the amount of emails sent. The coloring of the nodes is identical to the illustrative Figures in section 3.5.3: the custodians are black, the external domains are red, and the vertex representing the rest of the Enron corporation in the center is gray. Only communication between the custodians and/or the external domains is shown.

Apart from the Enron corpus, the entire tool has also been applied to the author's email collection. Distributed over four email accounts, it contains a total of 3,999 unique email addresses and 10,218 messages. Other than in the case of Enron, we have an ego-network, e.g. a network that focuses on just one single person. The only source for relational data between other individuals than the owner of the network are messages with multiple direct or copy recipients. Over the entire collection, there are an average of 1.67 recipients per message (which is significantly less than the Enron average of 7.16). In comparison to Enron, all recipients were modeled as particles in the simulation. The owner of the ego-network has been assigned a fixed position at the origin of the simulation space. Figure 65 shows a visualization of the entire social network as documented in the author's email. We can identify three large clusters, of which two are the work-related communications at two university institutes, and one is a private friends and family cluster. The coloring uses the clustering information to provide additional visual distinction and aid.



*Figure 63:* Visualization of the Enron social graph, displaying the communication occurring within the same cluster for a certain distance threshold  $d$ . The higher this threshold is chosen, the more global the communication that is displayed becomes (Image: F. Müller)



*Figure 64:* Visualization of the Enron social graph, displaying the communication occurring in week 32 of the year 2000. Direct mails ('TO') are indicated as gray arrows, copy emails ('CC') are indicated as blue arrows. The weights of the arrows correspond to the number of emails (Image: F. Müller)

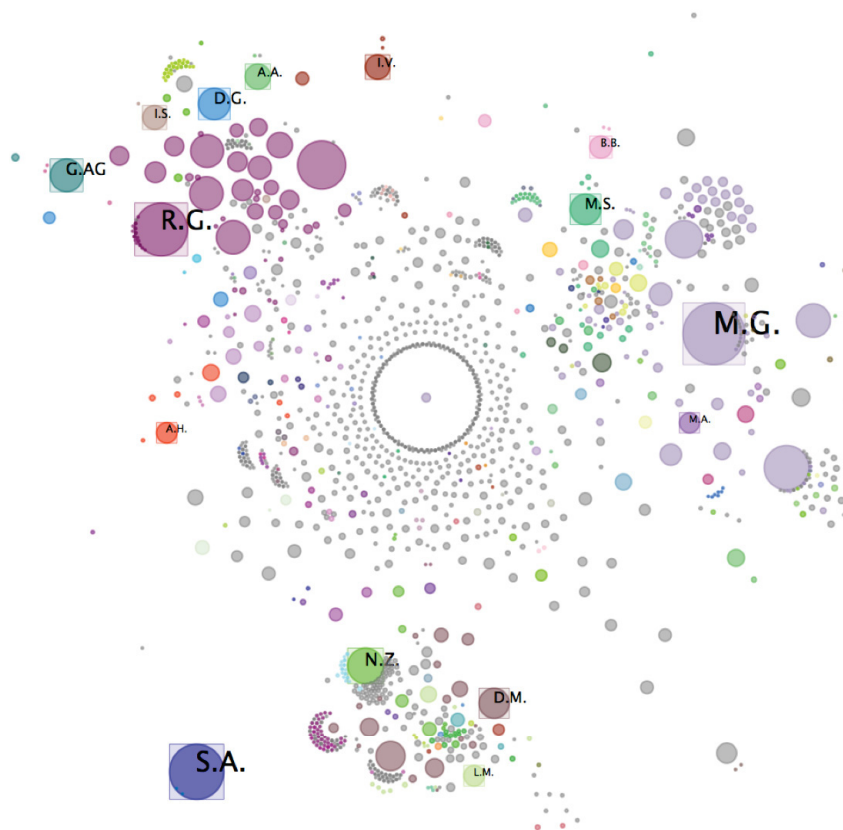
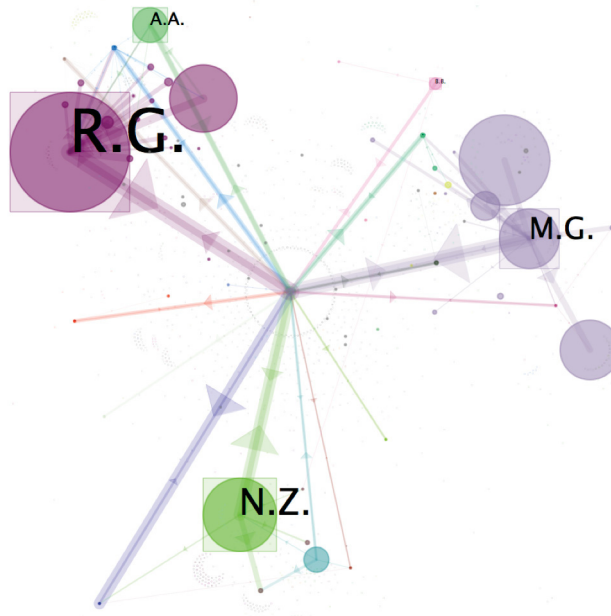


Figure 65: Social graph derived from several personal email addresses of the author. The owner of the ego-network is located in the center. Three large clusters can be identified, two are work-related and illustrate the network of two university institutes, one is private (Image: F. Müller)

Figure 66 shows a filtered view of the ego-network. While the absolute positions of the particles remain the same, the vertices and edges were weighted according to the frequency of the forwarding of emails. We have identified email forwards by performing a full-text search in the subject and the body of the emails, looking for a pre-defined set of strings (e.g. 'FWD:'). The figure shows who forwards information in the network, and in which direction it flows. It is a simple way to emphasize organizational activities.



*Figure 66:* Visualization of email forwarding in author's personal email. Forwarding indicates coordinative organizational activity (Image: F. Müller)



### 3.6 Summary

We have presented a method for analyzing relatively large social networks as documented by a set of email communications. As a specific task, we have tried to derive the organizational structure implicit in the network. Based on a comparison with the actual and known organizational structure, we can say that our approach has been successful in dynamically tracking several structural aspects. The result of the clustering and the derivation of the organizational structure should be seen as a useful hypothesis in the analysis of social networks. It aggregates millions of communications in a view that allows one to quickly develop a global perspective on the structural make of an organization.

This is also evident in our visualizations of the social networks. Both the Enron network and the ego-network of the author's email communication can be made comprehensible at one glance. The force-directed approach, together with the reaction-diffusion model, results in an enormous reduction of complexity. If we imagine such social networks as lists or as tables, we must admit that their understanding will require long study. If we succeed in obtaining a good visualization, however, the most important aspects of such networks can be seen immediately.



**Part IV**

## **Perception - Mixed Reality Interfaces**



Throughout the history of philosophy, various forms of so-called *skeptical arguments* have challenged our conception of reality. In a variant prominently shaped by Descartes, the relationship between our perception as we experience it and our perception as it is caused by an external world is under scrutiny.

If we do not trouble ourselves with skeptical thinking, we could describe a naive relationship between our perception and the world as follows: for every experience that is brought to us by our perception, there is one (or are several) real causes in the world around us, and the nature of our perception allows us to reliably conclude that it has resulted from a well-defined state of the external world. For example, when we eat an apple, then this experience – the taste in our mouth, the smell in our nose, the resistance of the apple to our jaw, the weight of the apple in our hand – is caused by us actually eating an apple. Our perception directly and truthfully informs us about the state of the apple and us in the world.

Descartes, in his *meditationes de prima philosophia* [152], has put forward a skeptical argument that can be referred to as *cartesian skepticism* or *skepticism of the outer world*. His thinking starts with a reflection of a well-known state where the borders of reality and illusion are all but clear: the state of dream. From his own experiences with dreams, he knows that in such a state, things appear as though they were real, although they are not. To go back to the example of the apple: if we dream of eating an apple, this dream can bring us a very vivid experience that may be indistinguishable (at the moment) from the real experience of eating an apple. It is only later, namely when we wake up, that we realize that we have not actually eaten an apple, but only dreamt of it. Even though retrospectively (and sometimes even in the moment), we can distinguish dreams from reality, Descartes concludes that it is in principle possible to perceive in a manner that is not caused by what it seems.

This has led him to formulate the hypothesis of the *evil genius*. If there were a god-like creature with complete control of the world and the people in it, and this creature were interested in deceiving the inhabitants of the world about how the world actually is, then she could do that by creating perceptions that are illusory. Essentially, such an argument results in the confession that our perception does not reliably inform us about our world, and that our knowledge of reality is limited in that it depends on the – in principle unintelligible – relationship between the real cause and the appearing cause of our experiences.

This little digression into the philosophy of mind is no end in itself, but serves to illustrate the conceptual environment in which the technologies presented in this chapter operate. In a more modern variant of cartesian skepticism, the god-like genius is replaced by technological components, and the resulting hypothesis is called the *brain in a vat* hypothesis. In this hypothesis, it is thought that we could in theory be nothing but brains cultured in a vat, connected to elaborate machinery that stimulates the synapses in our brains in a manner which makes it appear to us as if we were living in a real world and having real experiences. Brain in the vat scenarios have been popularized in fictional works such as *The*

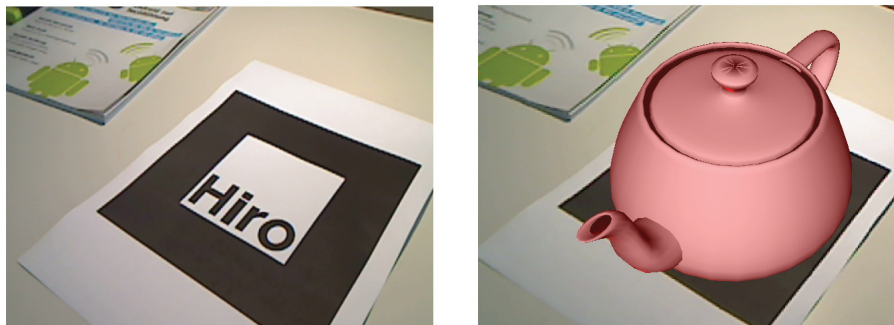
*Matrix* [153] or *Existenz* [154], to name just two. The philosophical debate on how to rebut such hypothesis (which is essential for our conception as beings who have knowledge of the world) is not our concern at this point. What we are interested in, however, is the ongoing endeavor to build technology that is capable of performing the tasks attributed to the evil genius or the brain-in-the-vat-machinery. More specifically, we will focus on one mode of experience and how it can be mimicked: the perception of space.

Computers are able to generate arbitrary representations of three-dimensional worlds (models) and render them on appropriate displays (views). Such worlds have first become prominent as *virtual worlds*. Sutherland, a pioneer in virtual reality (VR), has introduced the idea of “The Ultimate Display” [155]. He proposed that with the appropriate display, one could extend the range of possible perceptions without limitation: if we can perceive what we can model in computers, then our perception is no longer limited by the restrictions of our very real physical universe. In the past decades, virtual reality technologies have made rapid progress. The increasing computing power has allowed the creation of models of ever more complex virtual worlds, and the introduction of higher-resolution displays has made the perception of these worlds more realistic and, importantly, more immersive. In the context of virtual reality, *immersion* refers to the degree to which the experience of the virtual world is similar to our experience of the real world. When we experience a virtual world through a computer display sitting on a desk, we continue to perceive our real environment. If we use a head-mounted display, which occludes our view of our real environment, the virtual world is the only visual perception we have and is thus more immersive.

While VR focusses on creating entirely virtual experiences decoupled from our real environment, the concept of mixed reality tries to bring our real environment and any virtual environment into alignment. Instead of replacing our experience of the world with the experience of a virtual world, mixed reality aims at complementing the real world experience through the introduction of virtual elements. The different forms of mixed reality will be introduced in the next section. At this point, we should return to the initial skeptical thought experiment. In this scenario, we discussed the uncertainty of the relationship between our experience of the world and its true state *as a whole*. Virtual reality technology can be seen as an attempt to build technological equipment that is able to *implement* the mechanism on which skeptical arguments operate. An elaborate virtual reality environment can be seen as a *proof of concept* of the cartesian skeptical hypothesis. With mixed reality, we enter a different field. The aim is no longer to entirely substitute our experience of the real world with that of a virtual one, but to combine the two into a single experience. This experience aims at being coherent in that the two elements of it – the virtual and the real ones – should at first sight be indistinguishable.

## 4.1 Mixed Reality: Aligning Reality and Virtuality

The term *Mixed reality* refers to the combination of real and virtual elements in the sensual experience of a user. A mixed-reality environment is a combination of hardware and software that allows the user to have a sensual experience that originates from both real and virtual stimuli. In the following, we understand real and virtual as follows: a real stimulus can be considered as a stimulus that is caused by the very real object of which it is the appropriate stimulus (e.g. the visual impression of an apple when someone actually sees an apple), while a virtual stimulus can be considered as a stimulus of which the object causing it specifically aims at mimicking the real object it purports to be (e.g. the visual impression of an apple when someone sees an almost perfect projection of an apple). In other words, we assume the possibility of reliably informing experience (opposed to the skeptic) as given. The combination and seamless integration of real and virtual elements in one single experience extends the classes of experiences a user can have considerably. Figure 67 illustrates the concept of mixed reality. On the left side, we have an image of a desk surface with a magazine and a so-called marker (*Hiro*). The desk, the magazine and the marker are real objects captured by a camera. On the right side, we have a very similar image, only with the addition of a teapot. However, the teapot is not actually there – it is rendered into the camera image of the desk, the magazine and the marker. The rendering into the camera image is performed in a manner that makes it appear as though the teapot would stand on the desk surface.



*Figure 67:* Demonstration of the principle of mixed reality. On the left side, we have an image of a planar desk surface with a magazine and a special visual marker. On the right, this image is augmented through the introduction of a virtual teapot (Image: F. Müller)

Milgram [156] has introduced the *virtuality continuum* to classify various forms of mixed reality environments (see Figure 68). At the reality side of the continuum, we have an entirely real environment, namely the world as we perceive it without any technological inference. At the virtuality side of the

continuum, we have an environment that is entirely virtual – in non-technical terms, a dream in which our entire (vivid) experience stems from thought alone can serve as an example, while in technical terms, virtual-reality technology is traditionally concerned with providing virtual experiences. An example of an entirely virtual reality is Second Life, the multi-purpose online virtual world. In between, we can seamlessly vary the degree of reality (or virtuality). Environments which are mainly virtual and augmented by some real components are called augmented virtuality environments. An example are television studios in which only a desk and the speakers are real, while the rest of the environment (especially the background) is virtual. Finally, environments which are mainly real and augmented by some virtual components are called augmented reality environments. We will focus on these in the remainder of this part.

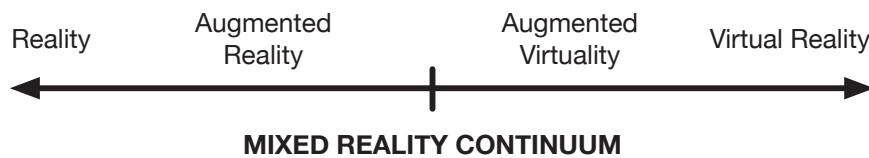


Figure 68: Mixed reality continuum (Image: F. Müller, after [156])

We should note that mixed reality technologies are, before all, display technologies. While reality and virtuality can be augmented through the path of all senses, the modulation of each of these senses requires different technologies. In the remainder of this section, we will follow this focus on visual technologies. Also, we will mainly consider the variant of augmented reality, where a mainly real environment is augmented by some virtual elements.



### 4.1.1 Technological Foundations

Augmented reality depends on a number of technologies. In order to gain a systematic overview, we will start by citing a definition of augmented reality that was given by Azuma [157]. He states that augmented reality is a display technology with the following properties:

1. It combines real and virtual
2. It is interactive in real-time
3. It is registered in 3-D

Since AR is a display technology, the combination of real and virtual refers to the combination of real images and virtual images. That these images (or rather, this image stream) must be interactive in real-time means that the system must react to what the user does – opposed to, say, the combination of virtual and real imagery in motion pictures as *special effects*, which are not interactive. The registration in three dimensions, finally, is probably the most challenging criterion. It means that the virtual contents of the image must have a specific spatial relation to the real contents of the image, and that this relation must be constant. For example, if we augment the image of a real surface with virtual elements, then the location of these virtual elements is defined as a location on the real surface – if we change our angle of view, or if we move the object of which the surface is augmented, then the augmentation must change accordingly (as if it were actually physically located on the surface). The first and the third requirement are best illustrated by describing the implications they have for the hardware of an AR system.

The first requirement (combination of the real and the virtual) implies that we need some image-generating technology (a display) that allows this combination of real and virtual elements. Among the more popular of these technologies are head-mounted displays (HMD). HMDs are small displays that are arranged in such a way that they can be worn on the head by a user. Two variants of HMD are usually distinguished. *Optical see-through* displays consist of a partly transparent surface mounted in front of the user's eyes, onto which the virtual elements are projected. *Video see-through* displays consist of two displays (typically LCD) mounted in front of the user's eyes, completely occluding what the user would actually see, on which one (or two, in the stereoscopic case) live video stream(s) captured from a camera (or two) located behind the displays (from the user's perspective) are displayed. Such a HMD is illustrated in Figure 69. Optical see-through displays provide the user with a very wide field of view, as well as an immediate visual perception of the environment (which, in consequence, is perfectly synchronized with the experience of the other senses). Video see-through displays, on the other hand, provide a better visual experience (the contrast on optical see-through displays is limited), and the control

of the registration of the virtual and real components is more extensive (for a detailed comparison of optical and video see-through displays, see [158]).

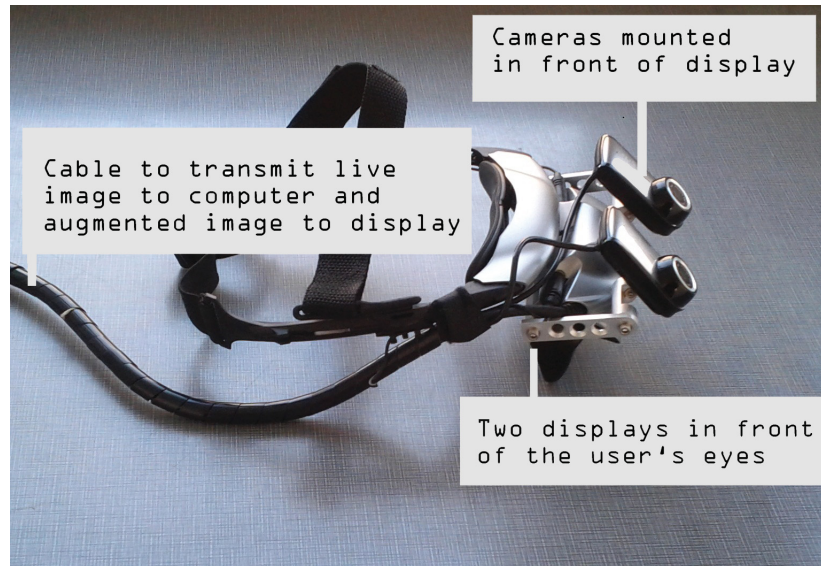
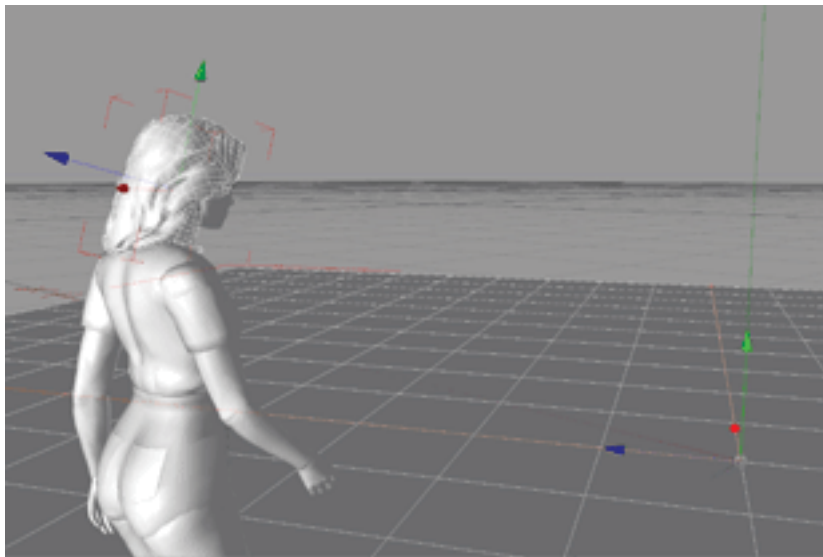


Figure 69: Photograph of a head-mounted video see-through display. The cameras in the front capture the user's field of view in real-time and transmit it to the computer through the cable. The computer processes the image and sends it back to the displays behind the cameras (Image: F. Müller)

The third requirement (registration in 3-D) implies that at any given point in time, the system must know the user's exact location in space as well as his field of view in order to register the virtual components (which are assigned real spatial properties such as location, orientation and scale) with the user's view. The user must therefore be *tracked*: her location in space (position) and the orientation of her view (orientation) must be monitored and continuously govern the display currently appropriate. This tracking of position and orientation is called 6-degrees-of-freedom (6-DOF) tracking. The six degrees of freedom are illustrated in Figure 70.

Various technologies exist to track a user. In principle, one can distinguish vision-based tracking methods and sensor-based tracking methods. In vision-based tracking, a live camera image is analyzed in order to determine the 6DOF of the user. Vision-based tracking can either be inside-out (IO) or outside-in (OI). In the IO case, the field of view of the camera is identical to (or at least in a known fixed relationship to) the field of view of the user. The user is localized within his environment by calculating the position of the camera within that environment. In the OI case, one or several cameras with different fields of view are used to determine the 6DOF of the user, which has to be in sight of the camera(s). The user typically wears markers that help the vision algorithm detect her body and pose reliably (see [159] for an example implementation).



*Figure 70:* Illustration of the tracking problem: both the location of the user in space and the orientation of her field of view must be tracked (Image: Lifeclipper2 project)

The use of artificial markers that are placed in the environment is a possibility in both IO and OI tracking. A popular and probably the most wide-spread vision-based IO tracking toolkit, the ARToolkit developed by the University of Washington [160], makes use of square fiducial markers with known dimensions to calculate the user pose relative to the camera image. The illustration in Figure 67 shows how the ARToolkit operates. While ARToolkit provides surprisingly accurate tracking and is easily extendable to write new AR applications, the use of fiducial markers limits the range of applications one can use it for. In more recent vision-based tracking approaches, natural features instead of fiducial markers are used in order to determine the camera pose [161] [162].

The most widely used sensor-based tracking approach makes use of the Global Positioning System (GPS) to track a users position. GPS localization is only available outdoors, and by default only offers limited accuracy in the range of several meters. Techniques exist to increase the accuracy of GPS, such as the use of virtual reference stations. Using such methods, the accuracy of longitude and latitude can be increase to one to two centimeters, and the accuracy of the height above sea reaches the range of 10 centimeters. It should be noted that GPS only provides three degrees of freedom. An AR system relying on GPS must include another sensor for determining the field of view of the user. A commonly employed variant is the gyroscopic sensor, which is able to deliver orientation information in very high accuracy and at high frequencies.

### 4.1.2 Evolution of Hardware Platforms

The technologies and devices that power augmented reality systems have seen a rapid development over the past decade. In a survey of various AR technologies, Papagiannakis provides a taxonomy for classifying AR technologies [163]. An extract is depicted in Table 11. The computing device contains the actual AR application and can be seen as an integrator for the other components: it obtains the tracking information, determines which virtual view to render and registers it with the live image and then sends it to the display. Some systems work only outdoors, some only indoors and some both outdoors and indoors. The restriction in the field of application is usually due to restrictions in tracking. GPS does not work in buildings, and OI visual tracking requires a camera setup which usually has only a limited range.

Computing Device	Mobile PC, Tablet PC, Ultra-Mobile PC, PDA, Phone
Location	Indoor, Outdoor, Indoor and Outdoor
Tracking	GPS, Visual IO, Visual OI, UWB, WLAN, Sensors
Display	Head-mounted, Handheld

Table 11: Taxonomy of AR technology components according to Papagiannakis [163]

Two factors have made it difficult for AR technologies to be adopted by a wide range of users: obstructive hardware and power consumption. A traditional hardware setup, as it has been employed in a multitude of projects [164] [165], consists of equipment that has to be worn in the form of a backpack, weighing several kilos, and of additional hardware for display and possibly interaction devices. Apart from being expensive and obstructive, it also requires a lot of power. Such systems can typically run only for a few hours, depending on the battery strength of the laptop.

More recently, AR research has seen the rise of a new platform that seems to solve many of these difficulties: smartphones. Today, smartphones have enough computing power to work on three-dimensional graphics in real-time. In addition, they have integrated cameras, a display and various sensors. Smartphones are also the first augmented reality platform to be available to a broad and growing user group. Among the research AR systems implemented on smartphones, the Studierstube tracker, a vision-based tracking engine, has received early attention [166]. First consumer AR applications for smartphones appeared in 2008 and 2009, such as Wikitude [167] and Layar [168]. With ever expanding possibilities, smartphones have brought about a breakthrough for the adoption of AR technologies and have led Bruce Sterling to announce the “dawn of the augmented reality industry” in 2009 [169].

After this introduction of mixed reality and its technological foundations, we would like to introduce two augmented reality projects in which the University of Basel’s Computer Science Department was involved in the past years. The first

one, Lifeclipper2, consisted in building a mobile augmented reality system for use in urban environments with a rather broad range of application scenarios. The second one, HUVis, focuses on architecture visualization and was developed for a client/server architecture with handheld thin clients responsible for capture and display and a server side responsible for the computationally intensive tracking and image generation.

## 4.2 Lifeclipper2: Staging Public Space

Lifeclipper2 (LC2) was an interdisciplinary AR research project and a collaboration of the Fachhochschule Nordwestschweiz (FHNW), i-art interactive AG and the University of Basel<sup>53</sup>. It was a continuation of the previous *Lifeclipper* project by the artist Jan Torpus. It investigated the potential of design and content of media performances in public spaces. The project team consists of researchers in the fields of art and design, computer science, architecture, and archeology. The main goal was the creation of a system that allows new experiences of city space including virtual spatial features such as buildings and parks. LC2 is able to visualize future developments, and could for example be employed in the exploration of city planning possibilities and in the evaluation of architecture competitions, making possible the near-life experience of concurring projects.



Figure 71: Live view from the Lifeclipper2 system (left) and user wearing the system on his back (right)

Lifeclipper2 focuses on a specific urban environment, namely the vicinity of the Novartis Campus, a pharmaceutical company's headquarters which is currently being re-built in the St. Johann quarter of Basel. The Lifeclipper2 system provided the user with four so-called 'scenarios'. Each scenario allowed users to experience a certain aspect of the environment, and was bound to specific locations in the quarter, as well as to a time frame.

An AR scenario specifies the virtual overlay (both static and interactive) that is associated with the given real-world environment. AR scenarios are spatially limited and explicitly dened. An AR scenario is a set of virtual 3D objects and multimedia elements that is bound to a specific area described by geographic coordinates. Inside an AR scenario, several behaviors and interaction possibilities are defined. The conception and the creation of AR scenario must be highly flexible and adaptive to support the broad range of wishes of the media

<sup>53</sup>The project was funded by the Swiss Federal Commission for Technology and Innovation (CTI) under grant number 8742.1 ESPP-ES. )

production team.

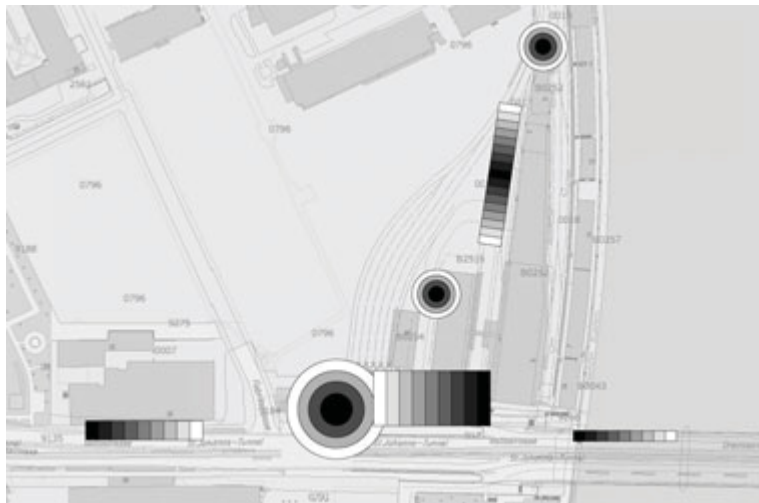
The first AR scenario consists of 4 areas located around a focal point of city development, where a new two-level bridge has just been completed over the river Rhine and the creation of the large-scale Novartis campus is in progress. The feature that most impressed the users was a virtual elevator. The elevator would take the user up into the sky, showing on a broad scale the future development of the area. Because the field of vision can be arbitrarily controlled, the fact that the users did not in fact leave the ground could be compensated well. The virtual elevator is also an example of a change in the degree of virtuality. When taking the elevator, the user does not actually leave the ground, so the imagery presented to her cannot reflect what she is actually seeing (namely, still the ground view), but must be entirely virtual. It could be said that at this given location, the AR system shifts towards a Virtual Reality system.

Another AR scenario we would like to mention is Archviz, short for architecture visualization. In this scenario, future spatial layouts of entire city areas can be visualized and experienced by the user. Large-scale development projects often cause inspired debate among citizens, and the capability of constructed models and still-image or video visualizations to capture the effective result of the development is often doubtful. Using AR systems to create an almost real-life experience of the development could truly alter the evaluation and perception of projects in the planning phase. The Archviz scenario allows the experience of the plans for the future development of one of the city ports. The evaluation of Archviz is ongoing, and upon completion will allow conclusions about the usability of AR technologies for city planning and architecture experience.

The LC2 system was demonstrated publicly during the official inauguration festivities of an underground transit highway (Nordtangente) at the location mentioned before. The AR scenario that had the broadest positive feedback was the visualization of a former Celtic settlement at the location of the research campus. It was developed in cooperation with archeologist from the university and the city, and has attracted great interest from both professional (archeology, tourism, city marketing) and leisure time users of the system. For the development of commercial AR systems or applications, the degree of interest for the various AR scenarios provides valuable information: in an urban context, the enhancement of already existing attractions with AR technology is promising.

#### **4.2.1 Technical System Implementation**

Figure 71 shows the LC2 system in action and its basic hardware layout, as well as a screenshot of a user view at runtime. It consists of a laptop computer, sensors, a head-mounted display and input devices. The computer hosts the AR software components and interfaces the sensors and the input devices. The sensors measure the relevant parameters of the system user, which are her



*Figure 72:* Map of the environment of the Lifeclipper2 project site, showing the hotspots for the different scenarios. The river Rhine flows upwards on the right and is crossed by a bridge visible in the bottom area, the Novartis Campus begins in the top area (Image: Lifeclipper2 project)

position in the world and the spatial orientation of her field of view. The input devices are able to capture user input and deliver it to the AR software components. The head-mounted display presents the overlay of the virtual and the real world to the user. A schematic depiction of the LC2 system information flow is given in Figure 73.



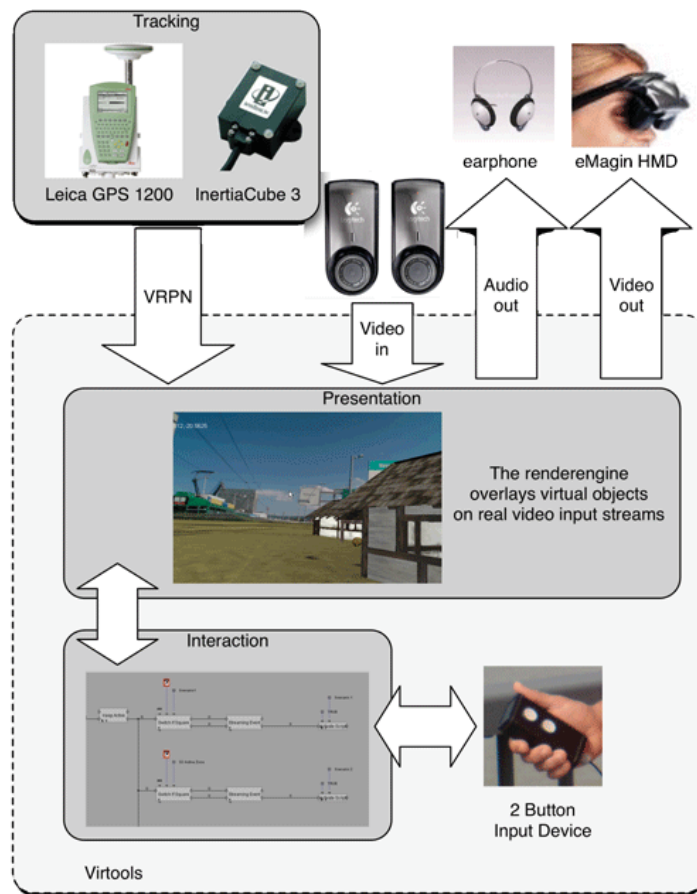


Figure 73: Schematic depiction of the Lifeclipper2 system in an information-flow perspective (Image: O. Koch)

Two cameras are attached to the front of the head-mounted display (HMD). They capture a live stream of the users field of view. The live stream is fed into the AR application, where it is augmented with virtual objects. The augmented live stream is then output via the presentation module and displayed on the two screens of the HMD. Two sensors measure the relevant parameters of the system user, which are her position and the spatial orientation of her field of view. The position is measured by a global navigation satellite system (GNSS), the Leica RX1250 differential GNSS [170]. It allows high precision position detection with a maximum inaccuracy of less than 0.01m. The system uses both GPS (Global Positioning System) and GLONASS (Global Satellite Navigation System) satellites and is able to improve accuracy through the use of reference stations. The GNSS consists of an antenna and a console. Both are mounted on the backpack. The update rate of the position sensor is variable and at most 5Hz. The user orientation is measured with an Intersense InertiaCube3 [171] that can measure the line of sight with an angular resolution of 0.03 RMS.

The update rate of the orientation information is  $180Hz$ . The sensors and the HMD are both connected to a laptop computer located on the backpack. The computer does not only interconnect all hardware devices, but also hosts the AR application software. LC2 uses both commercial and open source software packages. Among them are the model-software Virtools [172] and the Virtual Reality Peripheral Network (VRPN) protocol suite [173]. The LC2 system is designed for outdoor use and can be used for approximately two hours until the laptop batteries have to be recharged. When using several batteries consecutively, the system has been used (with short interruptions for battery change) for as long as 8 hours. The LC2 system has three main components: the tracking module, the presentation module and the interaction module. An architectural overview of the entire system, including the software modules, is depicted in Figure 74.

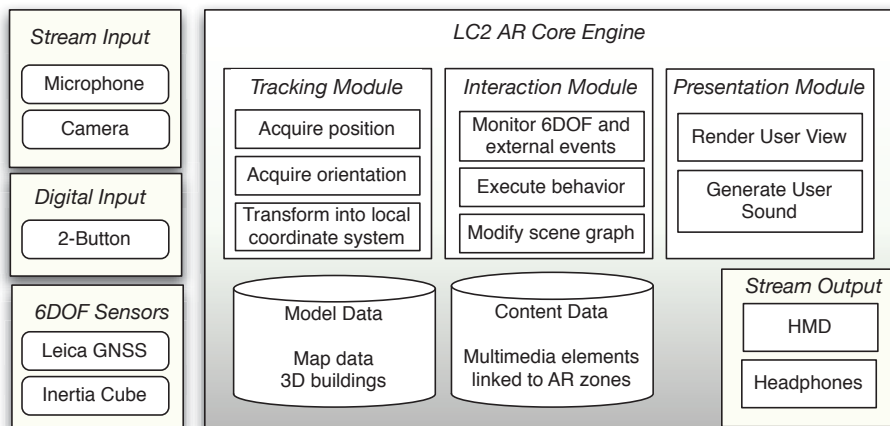


Figure 74: Architectural view of the Lifeclipper2 system (Image: F. Müller)

The tracking module mainly consists of several VRPN servers and clients. VRPN servers are software that capture the measurements of a sensor and transmit this information via the VRPN protocol. VRPN clients are software that receive the transmitted sensory information and make it available to other software components. In the tracking module, two VRPN servers capture the position and orientation sensors. Two VRPN clients within the core AR application receive the messages sent by these servers and make user position and orientation available to the model software.

The presentation module generates the user view. It captures the real view of the user line of sight as provided by the two cameras (live stream). Based on the users position and orientation (made available by the tracking module), it computes the virtual objects that the user faces given the current model. The live stream is calibrated (adaptation in orientation and size) and projected to a

mobile virtual screen plane within the model (backdrop). A virtual camera then renders the user view, that is, the field of view given the virtual objects and the live stream backdrop. The rendered stream is displayed on the HMD.

The interaction module is responsible for effecting changes in the virtual environment of the user. For this purpose, it captures user input via specific devices and monitors the relevant user parameters (position and orientation). Typical effects of said parameters are activation or deactivation of contents (visibility, audibility), change of render settings (different model views such as textured, wireframe), and shifts in the degree of virtuality or reality (from purely virtual to purely real, i.e. un-augmented, and between). The specific way in which the virtual environment is dynamically altered is called a behavior. LC2 has used several input devices, among them a simple two-button device and a mouse providing two buttons and a wheel.

The LC2 system renders user views based on the environment and virtual objects in real-time. The system clock is determined by the slowest component, which is the GNSS position update. It operates at a maximum of  $5Hz$ . At this speed, the system is well suitable for walking speeds. For use in faster motion, the system clock would have to be increased. The virtual objects are present in the system as three-dimensional model data. In the user view, the virtual objects are integrated into the world and thus have a location and spatial dimensions. The size of the objects ranges from around 50 centimeters to several tens of meters. LC2 is ideally suited for the experience of mixed real and virtual outdoor environments. The system allows the continuous passage from purely virtual and purely real visual experience and any degree of virtuality/reality in between them. Interaction between the application and the user is achieved in two ways: the location and viewpoint of the user determine the model-location of the rendered scene, and the user can trigger events by using a dedicated input device.

#### **4.2.2 User Reviews and Lesson Learnt**

In several phases of evaluation, the LC2 system was tested by users from the architecture and art community [23]. All users stated that LC2 was an overall positive and exciting experience, which could also be seen from their facial expressions when using the system. The users were immediately inspired in respect to their profession and came up with many new ideas on how to use a system like LC2 for their purposes.

For city planners, the possibility to navigate through the city with complete liberty (virtual elevator) was very revealing. It gave them the possibility to observe the application of their models in the real world and assess their accuracy. Also, they emphasized the utility of being able to experience the future city landscape at the very place where it will be built and with a very high immersion factor. As a negative point, they reported the absence of high precision position information in the shadow of buildings, which leads to inaccurate model align-

ment. In an urban environment, the limited high precision GPS coverage is a serious challenge. It may be necessary to include additional position information technologies and use a hybrid tracking system [174].

Our second main evaluation group were artists. For them, the possibility of more interactive elements like the virtual elevator is very important. Also, for them, the separation of content creation outside the AR system and content experience within the system is not desirable. They suggested an AR system in which the content of the scenarios would be created within the very system where they would be experienced. In other words, it should be possible to build artificial worlds using AR technology, and not conventional computer-based 3D modeling techniques.

The most important finding from the user reviews is that there are several professional communities with a strong interest in specific AR applications. Our AR system covers a wide range of outdoor applications. For user evaluation purposes, a limitation to context specific scenarios would have been helpful. Therefore, our system should be dynamically configurable for specific user groups. This is possible by the architectural organization in independent, cooperating modules. A first step towards such an implementation has been made by dynamically loading model data depending on the position of the user, which O. Koch has provided for the LC2 system [175]. In an extension of the two previous projects, the artist Jan Torpus has conducted further research in the *Lifeclipper3* project [176].

### 4.3 Hybrid Images for Architecture Visualization

We call an individual image that combines real and virtual elements a *hybrid image*. A hybrid image can be considered a 'still shot' of an augmented reality live stream and is thus a specific application of augmented reality technologies. The term hybrid image has been chosen to allow a distinction between the real-time and interactive nature of augmented reality and the static nature of a single hybrid image.

The application of augmented reality technology for still images has been inspired by architecture visualizations usually provided for large-scale construction projects which are considered highly significant both for their environment and their stakeholders. The goal of the *HUVis: Handheld Urban Visualization* project was to provide a platform for the generation of hybrid images in the field of architecture visualization. Planned buildings can be made comprehensible using either renderings of virtual models, real scaled-down physical reconstructions or renderings of real photographs with the virtual model fitted in. The generation of such hybrid images is a manual task that requires the work of a visualization expert. Figure 75 shows an example of such a visualization of a planned building for the University of Basel's *Campus 2020* project.



Figure 75: Hybrid visualization of a future university building in Basel, Switzerland (Image: Baudepartement Basel-Stadt)

The HUVis project aims at implementing and extending a workflow proposed by Snavely et al. in order to automate the visualization procedure. Arbitrary stakeholders (from city planners to local residents) should be able to generate

visualizations from any perspective they are interested in. They are to take a photograph from the (future) building ground and send it to a visualization server. Given the photo, the server estimates the parameters of the photo camera (position, orientation, opening angle etc.). Based on this estimate, it generates a (virtual) rendering of the planned building from the corresponding perspective. The user-provided photograph and the server-generated rendering can then be combined into one single hybrid image, showing the real scene with the augmentation of the virtual building. The main problem of such a setup is a precise estimation of camera parameters. Today, many mobile phones with cameras are equipped with GPS and compass. The photos they take can be tagged with this information. However, the GPS and compass of such devices do not provide sufficient accuracy for a visually sound result. They can only serve as a first estimate of the parameters. The structure from motion technique allows a more precise parameter estimation sufficient for visually sound results. In our workflow, we have used the software package that Snavely et. al. have made openly accessible.

#### **4.3.1 Overall Workflow**

The overall HUVis workflow is depicted in Figure 76. The user inputs a photograph of the environment where the future building will be constructed. Based on a previously computer point cloud of this environment, the server estimates the camera pose (6 degrees of freedom) of the input photography. The structure from motion method involved in this process is described in detail in section 4.3.2. Once the camera pose has been estimated, a view for the future building is generated. Given the approximate position of the sun (calculated from the date and time of the capture), a model of the future building and a model of its current environment, a rendering is created in which the future building is shown as it would appear from the place where the photograph has been taken from. The model of the current environment is used to calculate occlusion of the future building by current buildings. Once the virtual view has been generated, it can be combined with the input photography (image compositing). This process completes the desired output, the hybrid image, which can be sent to the user.

It should be noted that in principle, the camera estimation and the generation of the virtual view as well as the compositing of the photograph and the rendering are independent. The camera is currently performed through the structure from motion approach, but could well be replaced by some other mechanism. In the next section, we describe the structure from motion part of the workflow in greater detail.

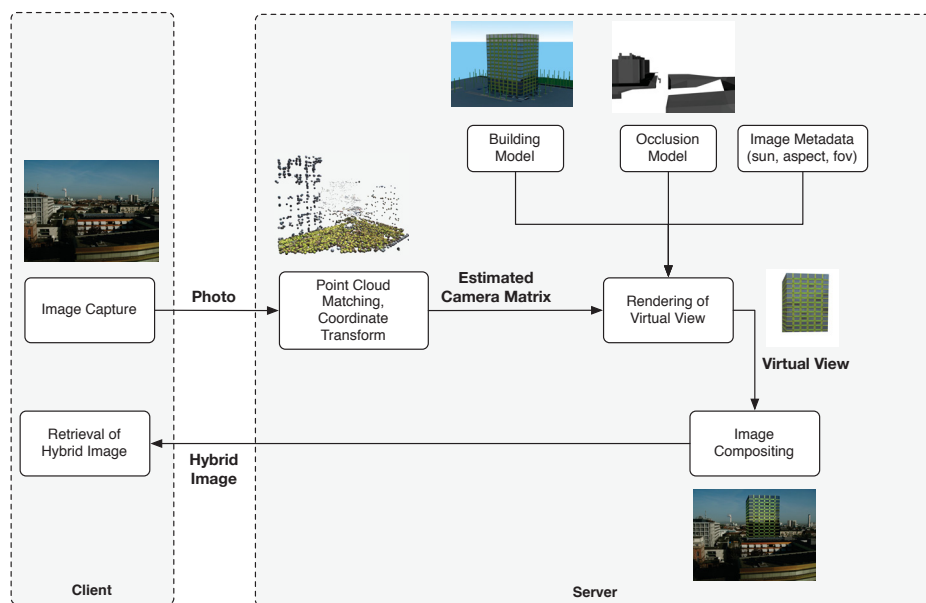


Figure 76: Current setup of the HUVis workflow. An input image (photo) is sent to the server, where the image is matched to an existing point cloud. The resulting camera estimation is transformed into the coordinate system of the model data. The rendering of the virtual view is based on the estimated camera, the model of the future building and its environment, as well as photo metadata. Finally, the virtual view and the input image are composited and sent back to the user (Image: F. Müller)

### 4.3.2 Structure From Motion and Point Cloud Matching

More formally, the mentioned structure from motion technique can be described as follows. Given a set of images  $I_1, I_2, \dots, I_n$ , image features  $f_{11}, f_{12}, \dots, f_{1n}$  are extracted for every image. A partial image input set is illustrated in Figure 77. The number of features depend on the scene (a blue sky will produce few, or zero features) and on the image resolution. We have found that typical urban scenes provide between several thousand (1024\*768 pixels) and several ten thousand (up to 3216\*2136 pixels) features. The features are SIFT (Scale Invariant Feature Transform) features as proposed by Lowe [177]. SIFT features are invariant to scale and to some extent robust regarding changes in lighting conditions. Each feature is expressed as a vector of length 128, and all the feature vectors of an image are the feature set of the image. Once all feature sets have been extracted, they are matched in order to find corresponding features between images. The decision whether a feature  $f_{11}$  in image  $I_1$  corresponds to one of the features of image  $I_2$  is based on measuring distances between features in their vector space. It should be noted that SIFT features are exclusively local – they are generated by measuring a 16 by 16 pixel area of an image, and do not contain global information of the image (for features including global context, see [178], [179]). Consider an image of a chessboard capturing several very similar features. Clearly, distance thresholding is not a good option here – the similarity between effectively distinct features makes selecting any single one of them very difficult. If one considers not only the closest feature, but also the second closest feature, it is possible to avoid making wrong decisions in situations where there are many similar features. Given a feature  $F_{11}$  in image  $I_1$  and features closest ( $f_{2i}$ ) and second closest ( $f_{2j}$ ) to it in  $I_2$ ,  $f_{2i}$  is only considered to correspond to  $f_{11}$  if the ratio between the distances  $f_{11} - f_{2i}$  and  $f_{11} - f_{2j}$  is below a threshold (0.36 seems to work fine [180]). The two decision criteria are illustrated in Figure 78.



Figure 77: Six images from the input set 'grid', showing different views of the same scene



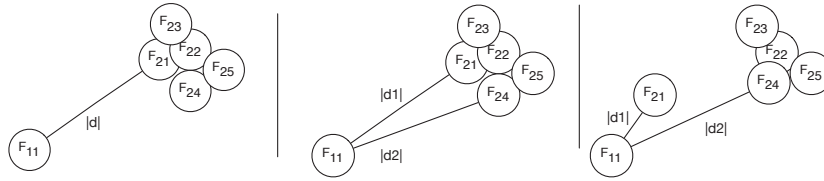


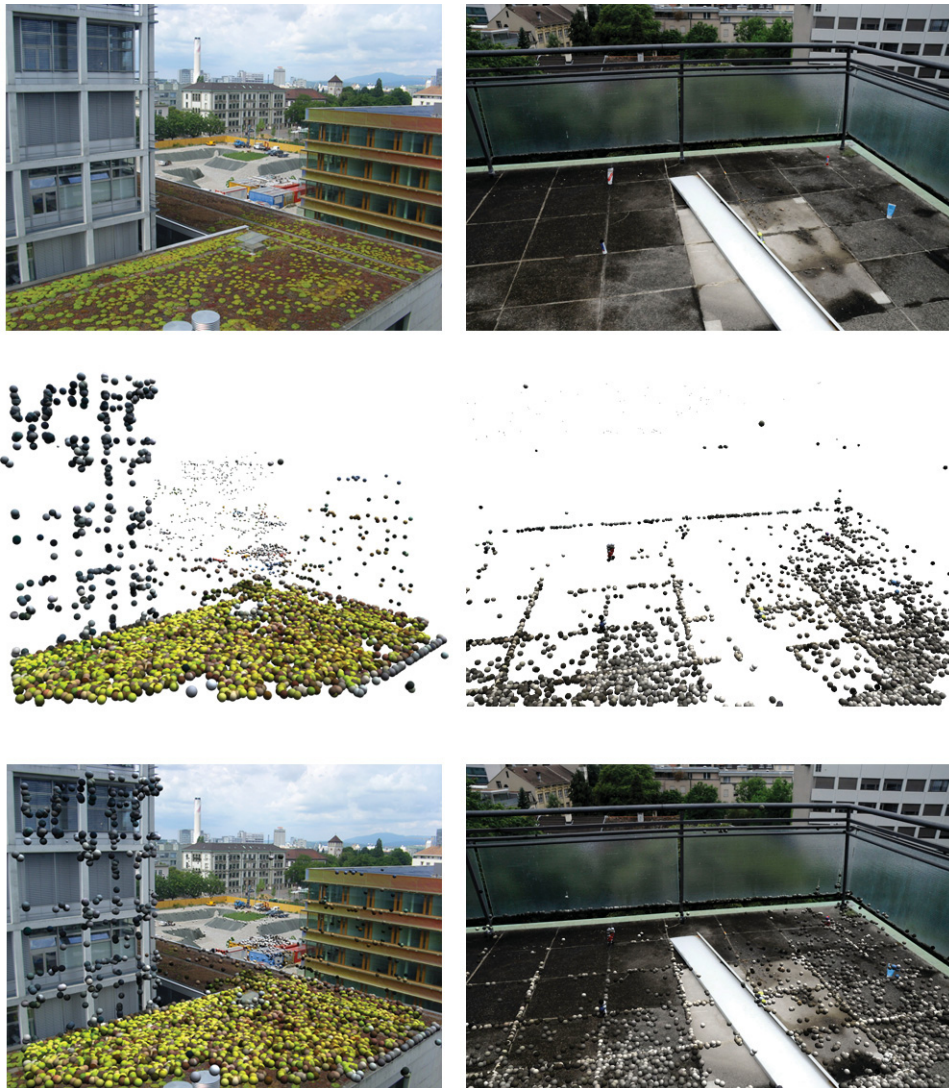
Figure 78: Decision criteria for matching feature  $f_{11}$  to feature  $f_{24}$ . Left: distance thresholding does not work well, because many features in  $F_2$  are similarly close to  $f_{11}$  – choosing any single one would be arbitrary. Middle: distance ratio between nearest and second-nearest neighbor. Since the closest is sufficiently closer than the second closest, it is not considered a match. Right: since the ratio between  $d_1$  and  $d_2$  is below the experimentally optimized threshold of 0.36,  $f_{21}$  is considered to match  $f_{11}$

The extracted features of all the images of the input set are the feature sets  $F_1, F_2, \dots, F_n$ . Every feature set has to be matched to every other set. Since the matching is performed in the high-dimensional space of the feature vectors, it is computationally expensive. Linear search between all feature sets is only viable for small sets and lower-dimension features. Lowe and others have proposed using approximative heuristics to perform the matching, such as approximate nearest neighbor search (ANN [181]). ANN provides an approximation to finding the nearest- and second-nearest neighbor in feature space and can be parametrized regarding its precision. It will be shown later that exact feature matching using auction algorithms is feasible even in high-dimensional space.

Once the feature sets have been matched, feature correspondences are known between all images. Given that enough corresponding features have been found, the scene geometry can be successfully inferred. The reconstruction consists of the  $n$  estimated cameras of the input image set and a number of points that represent features that are visible in several images and have been matched between them (called the *point cloud*). In order to determine the accuracy of the reconstruction, it can be rendered from one of the input image camera perspectives and overlaid over the input image. Figure 79 contains input images, visualizations of the calculated point clouds, and overlays. While the reconstruction contains the information provided by the known input images, it can also be used to estimate camera parameters for new images that show the same scene from a previously unknown perspective. For this purpose, a workflow similar to the original reconstruction is performed. Features are extracted from the new image, and are matched to the features of the images in the reconstruction.

Because linear search does not scale well with higher-order feature vectors, the ANN heuristic is employed. We have used auction-based matching to determine the scalability of exact feature set matching for structure from motion applications. We have used two scenes for our tests. They were captured using

two different high-resolution digital cameras<sup>54</sup>. From the original images, several scaled-down versions were created. The scaling down results in a reduction of the number of features that are extracted from the images. While using an input set with a high number of features is computationally more expensive, it allows a more detailed and more accurate reconstruction.



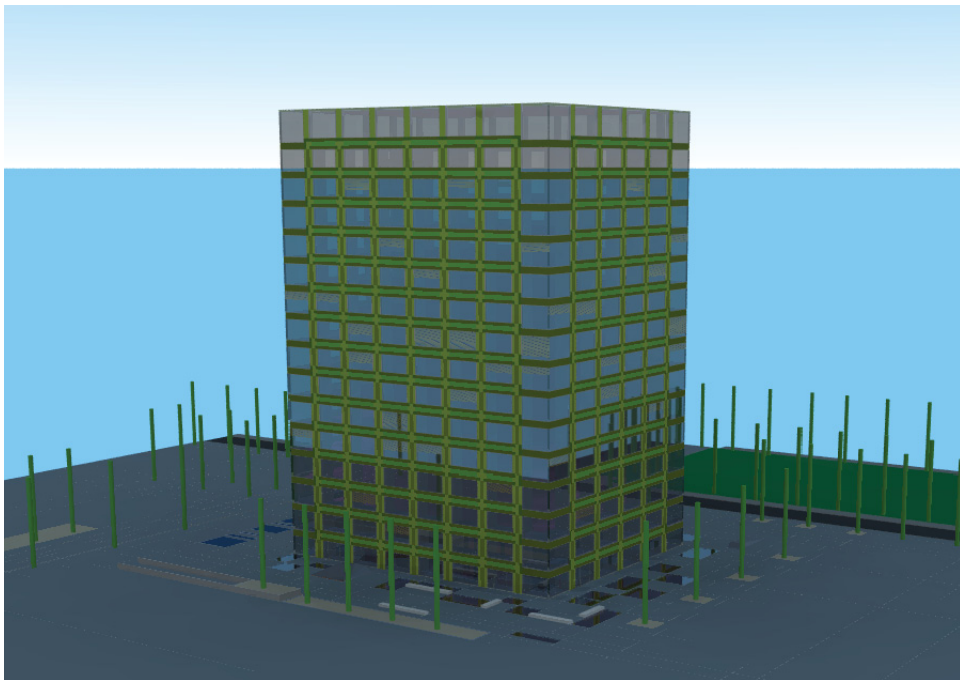
*Figure 79:* A wide-area (left) and a narrow-area (right) scene as seen from a specific perspective. Top: input photograph, middle: point cloud visualization from the estimated camera perspective of the input photograph, bottom: overlay of the input photograph and the point cloud visualization. The size of the points in the point cloud visualization does not reflect the size of the actual features.

---

<sup>54</sup>BenQ i750 and Nikon D300

### 4.3.3 Model and Virtual View

In order to generate the virtual view, several components are required: a three-dimensional model of the future building as well as the current environment, metadata of the photography, as well as additional materials for obtaining a visually pleasing rendering. Figure 80 shows a rendering of the model of the future building. Given this model, a view of the building from any possible perspective can be created.



*Figure 80:* Model of the new building (Image: F. Müller, based on LSZ model provided by Baudepartement Basel-Stadt)

In addition to the model of the actual building in question, the model of the immediate environment is required. The buildings in the immediate vicinity are important because they partly occlude the new building – in any view, we must take them into account and only render the parts of the new building that would not be hidden by the existing buildings. Figure 81 shows a hybrid image combining a photograph of the environment, and the future building as well as some surrounding buildings as partially transparent gray volumes. The environment model is derived from a 3D city model which was made available by the canton of Basel-Stadt. It should be noted that it does not (as of yet) contain model information of trees and non-building structures such as wooden walls for construction sites. The perspectives we have chosen for the project have in common that they provide views in which such structures would not

occlude the future building.



*Figure 81:* Hybrid image showing environment of the new building. The new building, as well as surrounding buildings that may from some perspectives occlude the future building, are rendered as partially transparent gray volumes (Image: F. Müller)

Several metadata of the photograph are used in the creation of the virtual view. The size of the input images determines the size of the rendering. The aspect ratio as well as the aperture of the physical camera that was used to take the photograph are used to parametrize the virtual camera of the rendering. In addition, the date and time as well as the geographical location of the capture are used to calculate the position of the sun, which is used as the position of the light source in the rendering. The virtual view is generated using Povray [182], a ray-tracing engine that operates on textual scene descriptions and is thus ideally suited for a client-server environment.

Figure 82 shows a hybrid image that combines a photograph of the scene with the calculated point cloud (blue spheres) as well as the model of the future building. While we have used a three-dimensional model of the environment to calculate occlusion, one can see from the figure that in principle, it would be possible to use the point cloud itself (which contains the neighboring buildings) to calculate occlusion (methods exist to calculate dense and thus actually occluding meshes from pointclouds [183]).

Once the virtual view has been generated, the resulting rendering is combined with the input photograph. We have used the ImageMagick toolsuite to

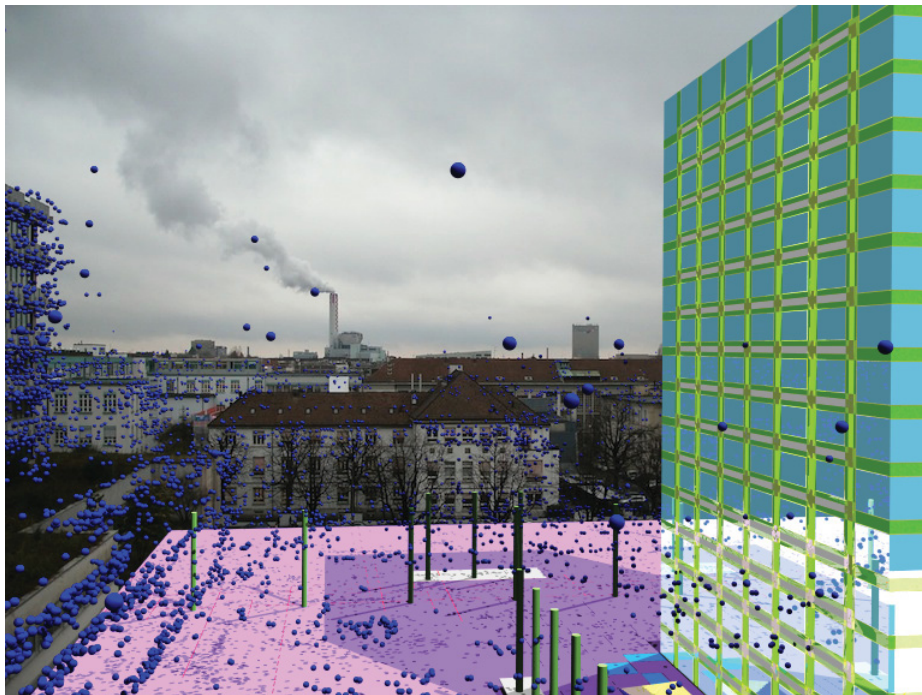


Figure 82: Overlay of point cloud (blue), model and photograph (Image: F. Müller)

composite the final image. For a more detailed description of the workflow including the exact specification of the tools, the reader is referred to the Master thesis of Tobias Denzler [25].

#### 4.3.4 Hybrid Image Results

In our evaluation of the HUVis workflow, we have come to the conclusion that in principle, it achieves the goal it has been implemented for, namely to provide a platform for architecture visualization that would allow lay people to produce their own visualizations of future projects with very little effort. That said, we consider two issues to require further investigation for the platform to be truly successful.

First, the camera pose estimation via the structure from motion workflow is quite expensive in terms of set-up and calibration. Consider the environment depicted in the various illustrations of the HUVis project. It is noticeable that the images are always taken from a quite similar perspective, namely looking north-west. The reason is that in order to be practicable, we have worked with point clouds that are based on photographs from similar perspectives. If one were to cover a view on the scene from a 360-degree circle around it, the resulting point cloud would be based on a very large collection of images (we estimate it to be in the thousands). While such large point clouds are in principle possible, they

are very costly in terms of computation, as has been shown in section 4.3.2. And even if one were to work with several smaller point clouds (e.g. based on 8 or even more ordinal directions), the individual point clouds would still have to be created, which is time-consuming. In addition to carefully taking the photographs (possibly under several lighting conditions), a good translation between the coordinate system of the point cloud (which is created arbitrarily by the structure from motion process) and the coordinates of the real city space must be identified. We have achieved this by measuring reference points in real space which are identifiable as individual points in the point cloud, and then calculating a transformation between the point cloud and real city space using a Seven-Parameter transform (see [184] for suitable software).



*Figure 83:* Hybrid image showing the new building in its environment, using an omnidirectional panorama as the light source for the scene (Image: F. Müller)

Another issue was the photorealism of the resulting hybrid image. Using artificial light sources and the texture information available from the three-dimensional model, we have found the results unpleasing. Figure 82 shows the visual appearance of the future building without any enhancements. The rendering in Figure 83, on the other hand, shows the resulting hybrid image when using an image-based light source, a so-called *light probe*. A light-probe is an omnidirectional panoramic photograph that can be used as the (spherical) light source of a scene [185]. The advantage of using such a technique is the realistic reflection on the building surface, which is evident in the illustration. A

disadvantage is the relatively high cost of implementing such lighting. One can either construct a custom omnidirectional panorama, or use a generic panorama with a comparable setting. In either way, one must have several such light probes to reflect different lighting conditions (be it due to the time of day or the weather conditions).

#### **4.4 Summary**

Mixed reality technologies enable us to detect, reproduce and alter almost arbitrary perceptual situations. Since they allow the seamless integration of virtual and real content into one single experience, they do not only enhance the possibilities of computer interfaces, but our real environment.

The Lifeclipper2 and the HUVis demonstrate not only the potential of mixed reality technologies, but also highlight how such technologies are more and more available to every-day users of computers. In Lifeclipper2, an elaborate hardware setup is required in order to provide the user experience. We have built one single system, which can be used by one single user at a time. The cost of the application would be prohibitive for a broad introduction, and the physical limitations imposed on the user by the wearable hardware conflict with the aim of providing a natural perceptual situation. The HUVis project, only requires the user to provide a photograph taken with her own device. While the application can not be considered augmented reality strictly speaking (it is not interactive in real-time), the platform for which it was designed – smartphones – are in principal capable of real-time experiences.

In the coming years, interfaces that are integrated into everyday-perception will increasingly become important. The basic technological foundations have been laid in the past decade, and development is ongoing rapidly. With such interfaces, electronic memory infrastructure will be able to bridge the gap between external technological artifacts and internal personal experience.





**Part V**  
**Conclusions**



## 5.1 Electronic Memory: Functions and Utility

Electronic memory has been introduced as an electronic equivalent of our biological memories. The notion has been conceived against the background of lifelogging – the continuous and image-centric capture of impressions of one's life, and the harvesting of this capture. It is characterized by a functional view – functions that we can observe in our biological memories, such as specific retrieval, recollection, or reminiscence, are mapped to the electronic world. The result is a proposed extension of our biological capabilities with electronic means. It is in this sense that the notion of electronic memory is most ambitious: it enhances our biological memory through the vast store of our digital assets and the computational intelligence operating on it.

These digital assets cover a broad range. We have traditional documents such as photographs, text documents, spreadsheets and notes that we author ourselves. We have various collections of content in the form of books, audio and video. We have data from all the applications we use, all the sites we visit in the World Wide Web, and all the communication we have through channels operated by computers. And these, if we use a smartphone, include nearly everything except for face to face conversations – on which, however, such a device may eavesdrop. All in all, our digital assets are increasing, and the value they actually and potentially have is increasing as well.

The fundamental question is: what are we to do with all this data? Even if all we want to do is just keep it, we face several challenges as documented through the case study PEVIAR. And considering the amount and the potential hidden value of all this data, merely organizing and preserving it will not suffice. The data's potential value is based on the implicit information contained in it. The case study about the Enron email corpus has shown that the value of a corpus of social relational data is far greater than that of the sum of all individual messages. How do we view the conversations that take place and the social network that are constituted? We have indicated that in order to unveil them, we must be able to retrieve memory from the social graph, developing metaphors that allow us to abstract from individual relations and look at the overall structure. These metaphors function as filters – they reduce the vast amount of data on which they operate to aggregations that are intelligible and usable.

Finally, the use of the information contained in our digital shadows is decisively shaped by the interface we use to retrieve it. The case study on Mixed Reality shows how display technologies can help bridge the gap between merely accessing a document in a digital archive and vividly recalling memories from an electronic corpus. We largely obtain our sense of reality from our perception, and the duality of perceived and actual reality has been briefly explained in section IV. The ability to actively and almost arbitrarily shape our perception allows the seamless and complete integration of computer interfaces into what we perceive as our natural environment. By that, mixed reality technologies are

able to considerably extend the space of possible experiences.

The functional view of electronic memory introduced in section 1.2 assumes that electronic memory is to serve as an extension of our biological memory. The functional requirements reflect this in that they try to model and parallel our natural abilities. In an infrastructure-centric perspective, the three case studies presented in this thesis highlight different aspects of this infrastructure, investigating problems, possible solutions and prospects of future possibilities. We would like to briefly reiterate the meaning of each of the case studies for electronic memory infrastructure. *Preservation*, as investigated and offered by Peviar, is an indispensable component of any memory infrastructure. Our intuitive notion of abundant and reliable storage as trivial achievement of electronic storage technologies will continue to be challenged by the mentioned problems of digital preservation. *Evaluation*, employed exemplarily on a corporate social graph, plays a key role in coming to terms with the sheer amount of data available. The ease with which we structure perceptual data into information as biological beings can help to inspire new algorithms and procedures to identify what is relevant and what not. *Perception*, finally, and its engineering, will integrate our real and our virtual environments in one single space. While temporal and spatial distances will remain instantaneously insuperable, the speed at which the Internet interconnects computers on the entire planet will enter the domain of our experienced world.

In the remainder of this conclusion, we would like to contribute the following components. First, we look at how the concept of electronic memory is exemplarily shaping the details of some state-of-the-art applications developed today. Then, we try to illustrate one of the major conceptual components of this thesis – the *metaverse archive*. Finally, we outline some impacts of the presence of electronic memory infrastructure both at an individual and social level.

## 5.2 Electronic Memory: First Applications

It has been stated that our digital shadow has seen rapid growth in the past years. Individual capture activities such as taking pictures and creating documents have a tradition that predates the age of electronic information processing. What is new, however, is that all this information is becoming more and more networked.

As of now, innumerable services allow us to share our content with friends and the general public. Some of these tools are very specific, such as a photo sharing service, some are multi-purpose and cross-media, such as Facebook. We exemplarily choose Facebook to illustrate how such content sites are becoming hubs of digital information from and about us. While much of our digital shadow still resides on our individual computers, the online coverage of our lives is increasing rapidly. Social networking software such as Facebook can be seen from two perspectives. In a networked view, it allows the shaping of communities and details the various roles and cultures in such communities. In a user-centric view, social networking software can also be seen as an infrastructure for a lifelog. While the degree to which people use social networking sites to document their lives varies, all users have in common that no matter what the granularity, their life is documented to some extent. If we call a Facebook profile lifelog data, we can conceive any interface Facebooks (or third parties) provide as a candidate for an electronic memory application. And the degree to which such applications try to evaluate our data in a memory-centric perspective is already considerable.

This is well documented through the introduction of the Facebook *Timeline*<sup>55</sup>. The timeline is a manner in which the user's profile, based on her contributions to the social network, is presented. Instead of showing a summary of the most important details of a profile (such as where someone is from, in what profession someone is active, what someone is interested in, etc.), the Timeline aggregates all the content a user creates over time on a temporal axis<sup>56</sup>. The bottom of the axis represents the birth of the user (or her joining of Facebook, although backdating of content is possible), while the top of the axis represents the current time (or, presumably, the time of death). The user can navigate the axis in both directions, and details for every entry are available upon click.

The Timeline interface is a first step towards the development of useful e-memory applications. It operates on a user's entire data collection and tries to aggregate an interface that includes all relevant information about someone, not just a few aspects. So far, the interface is not specific regarding functional electronic memory requirements. Timeline can be used to reminisce or recollect, retrieve or reflect, but does not seem to favor one function over the other. But as such interfaces are developed further, we can expect to see more specific applications. And as more and more data of and about us is not only

---

<sup>55</sup>Presented at *f8* in September 2011 [186]

<sup>56</sup>The idea of using a timeline-centric representation as a desktop computing metaphor has been introduced with the concept of the *lifestream* by Freeman and Gelernter [150].

available to individual sites such as Facebook, but can be made available to any application we might be interested in, the awareness of electronic memory infrastructure will reach the broad public.

### 5.3 The Metaverse Archive

We have introduced the concept of the metaverse archive, and as stated before, the metaverse archive should not be considered as an implementation of an archive or to provide a specific archiving service. Rather, it is a concept that helps us understand the function and meaning of archives in a world of electronic memory. In this section, we would like to clarify this understanding.

The archiving of information is an activity that has been performed over millennia. Considering only analogue artifacts, one could say that archiving mainly consists in safely keeping said artifacts in an environment that minimizes the effects of physical decay. It has been stated that for many traditional items in archives, keeping them locked away and accessing them only very rarely has proven to be a successful method of preservation. The first digital revolution – the development of digital text – has introduced an important change to the archive. The information digital artifacts represent is no longer inevitably linked to the physical decay of the carrier, since it can be perfectly copied onto one or several new carriers. The resulting independence from any specific material carrier allows – at least theoretically – the perfect preservation of information for unlimited timespans. It seems, however, that some of the consequences of this remarkable development are becoming truly evident only now, after having witnessed a second digital revolution: the development of universal computers and their accompanying electronic storage that can store arbitrary digital data. One of the reasons seems to be the effort required to reproduce digital information. For centuries, digital text was reproduced by hand. The introduction of the printing press greatly reduced this effort. But the introduction of electronic storage has made this reproduction virtually effortless – additional copies of documents can be created at unprecedented speed and cost.

But this is not the only novelty electronic storage has introduced. New problems have arisen, and in the traditional preservation perspective of archives, hardware and software obsolescence are among the most prominent ones. The PEVIAR study has shown what measures are required to preserve digital information in a manner comparable to traditional archiving practices. But we believe that the traditional notion of the archive will further be challenged by the progress of the digital age.

An archive can be considered as a spatially and conceptually delimited entity. It is conceptually delimited because it is a place where valuable information is kept in the long term that would not have a place elsewhere. It is spatially delimited because the infrastructure to preserve information in the long term is costly and therefore limited. This is contrasted by a trend that can be observed throughout the history of computing: the trend towards abstraction from concrete infrastructure, both at the hardware and software level. Take the example of storage. While data is stored on a *physical carrier*, the concept of *logical volumes* allows storage on infrastructure that may span across several physical carriers. And going further and thinking about *cloud infrastructure*, it

becomes clear that we perform storage operations in a logical space: we store and retrieve documents to and from the cloud, and the hardware on which the cloud service is implemented bears little or no significance for us<sup>57</sup>. Apart from the hardware implementation being transparent, the cloud service is also ubiquitous. While it does have a certain physical location, the speed at which data travels makes it available in nearly the same manner all over the world. And it is the same speed that will eventually make the distinction between information that is archived and information that is live irrelevant. Speaking in terms of distances, the archive is no longer far away both spatially and temporally, but rather available instantly.

If we try to break this down to more familiar terms of archiving and to the manner in which current digital archives operate, we can illustrate the functions of a metaverse archive in terms of the Open Archive Information Systems (OAIS) standard. The OAIS is a reference framework for digital archives originally developed by the Consultative Committee for Space Data Systems (CCSDS) and later finalized as an ISO standard [29].

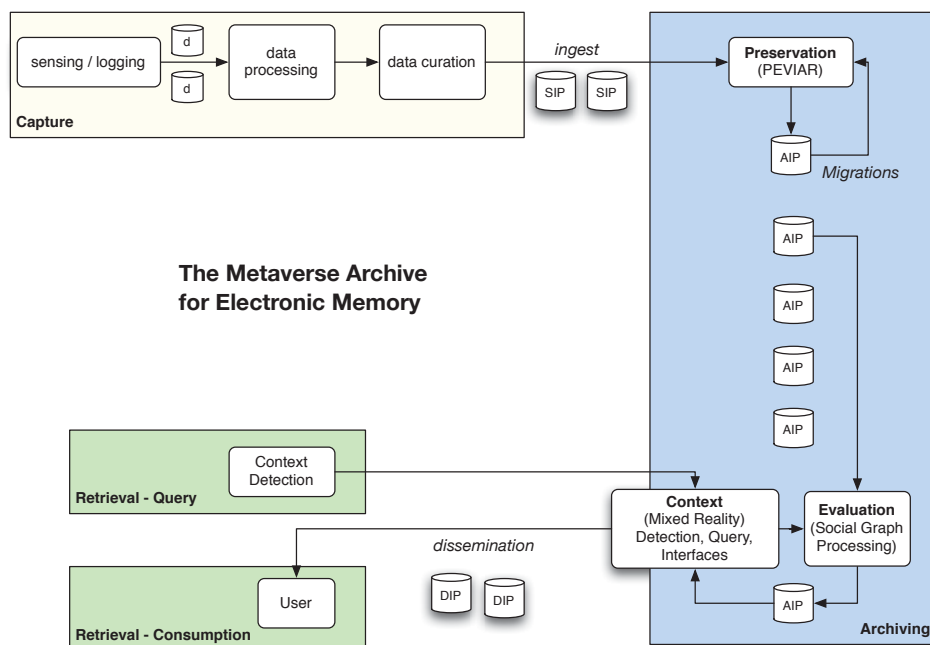


Figure 84: Schematic overview of the metaverse archive

While the OAIS has been briefly mentioned as a foundation of the preservation model introduced in part II, we would like to briefly explain some of the core

<sup>57</sup>That is not to say that the hardware implementation is irrelevant – on the contrary. However, we can consider it as a completely separate problem that is addressed not by the party requesting, but the party offering the cloud service.



concepts. The OAIS provides a functional model for digital archives in which the important stakeholders and technical components and functions are defined. The need for a coherent view on digital archiving was noticed early by Space Federations from around the world, which traditionally operate on large quantities of data gathered from their missions. Apart from the definition of stakeholders and components, the OAIS defines several workflows for digital archives. At the most general level, the workflow of a digital archive according to OAIS can be described as follows. First, some information is generated by a producer and prepared for admission to the archive. The preparation consists in creating a *submission information package* (SIP), which contains the information itself together with required metadata and is acceptable for archive ingestions. Once the SIP has been ingested into the archive, some software operating on the archive will derive an *archival information package* (AIP) from it. This AIP is suitable for the long-term preservation (subject to whatever measures the archive software may take on its AIP) within the archive. The information is now being preserved, and as soon as the archive receives an appropriate request for an archived item, it will prepare a *dissemination information package* (DIP), which packages the desired information in a manner that is suitable for consumption. The DIP is transmitted to the requesting party, and the workflow has been completed. While this is only a simplified depiction of the workflow, it should serve sufficiently to demonstrate the concept of the metaverse archive.

In the case of the metaverse archive, the principle workflow stays the same. Someone generates information and submits it to the archive, the archive ingests it and preserves it, and at any time in the future, it is able to transmit it to any party requesting it. However, for the metaverse archive as depicted in Figure 84, we *specify* a certain manner in which these actions occur. Under the term *capture*, we understand the automatic and continuous ingestion of information into one's personal archive. The main idea here is that submission of information is not explicit, but rather guided by policies the user defines. Such policies govern what sort of data should be continuously ingested, how it should be processed, and in what form it should be persisted.

Once information has been ingested into the archive, a new mechanism can now operate on the AIP, namely what we call *evaluation*. Even if we say that an information has been archived, we should not consider it a static resource. This has two sides. From an archivist's perspective, it is imperative that the packages in an archive remain intact, i.e. they must never be changed (or, if so, these changes must be documented). But then again, the idea of evaluation is that the information in the archive evolves – new insights are gained, and they must be reflected in the archive. Thus, we propose that the evaluation leads to the introduction of new AIP *from within the archive*. Evaluation should also have the ability to delete packages that were submitted from within the archive.

Finally, we have a mechanism for dissemination which considers context. Context detection, of course, happens outside the archive. What we mean by the context mechanism is that it must be able to support contextualized queries,

locate the appropriate packages, and then disseminate them in a manner which is appropriate for consumption. If we consider that mixed reality technologies play a crucial role in the consumption of this content, then the archive must support such interfaces in its exporting of information packages.

In summary, the metaverse archive leans on the established OAIS standard and attempts to clarify the required infrastructure for powerful electronic memory applications. It should serve as a bigger picture for researchers and developers who work on topics that are loosely or densely related to electronic memory. As we have seen, the involved technologies cover a broad range. The choice of an archiving framework to illustrate electronic memory, as has been stated, is due to our consideration that electronic memory is best seen as a powerful approach to personal information systems. Since preservation is a *sine qua non* of such systems, a digital archive model as the provide of this base service seems like a good reference framework.

## 5.4 Privacy, Control, and Transparent Citizens

The introduction of social media and exploding corpora of social data promise to bring new forms of social cooperation and knowledge. It has also brought about a complex discussion about privacy and control. We would like to discuss some of the implications.

Increasing spread and openness of social data also poses challenges. It was noted that in a manner, social computing brings us closer to the vision of the WWW as a global village. But in contrast to a real village, where one can leave for *another* village if life in the community becomes unpleasant, such an escape route is not easily found in the global village – there is only *one* village. As our real and virtual identities converge – on Facebook, users usually go by their names, not by a pseudonym – one’s identity is more and more inescapable. And since one is socially embedded, what this identity looks like is not entirely up to one self. As social media document more and more of our lives, they necessarily include documentations about the lives of others, namely the individuals we interact with. If I delete digital records about an episode I find embarrassing, it may well be that this episode is documented (in whatever form) by someone else. When considering, for example, an individual photograph that was taken at an over-enthusiastic moment, this may not seem like much, and it probably is not. As more and more people use social media, more and more embarrassing aspects of their lives are uncovered, with the result that ultimately, one’s individual embarrassments disappear in the noise of our collective embarrassments.

However, the fact that individuals become more and more transparent through their digital assets is well-documented for a case where these assets are beyond the control of the individual. In 2006, the European Union published what is informally called the *Data Retention Directive* [187]. It obligates member states to implement legislation regarding the mandatory retention of communication data by providers of public communication services, such as telephone network operators or internet service providers. An according federal law was introduced in Germany in 2007, which obligated communication service providers to retain customer data for at least six and at most seven months<sup>58</sup>. A politician successfully filed a suit to obtain the data stored about him by his telephone service provider. In collaboration with the newspaper *Die Zeit*, he published a subset of the data [188]. For a period of several months, one can follow the politician in an interactive map interface. Apart from his whereabouts and travels, all his phonecalls and short messages are documented (although the content, as well as the number of the respondents, are not disclosed). The author of the article has used public information sources such as Twitter or political blogs to infer the activity on any given day. The result is a surprisingly complete profile of the politician, even without the inclusion of the data on the other communication

---

<sup>58</sup>The law was nullified by the constitutional court in March 2010, making a new legislative approach to the implementation of the EU directive necessary.

partners.

As individuals and as a society, we will be confronted with the fact that our lives will be documented at a level of detail previously unthinkable. More and more tools allow the analysis and evaluation of this data. If the vision of the MyLifeBits project at Microsoft Research, namely to capture an individual's entire life and making it available for *total recall* [7], becomes reality, then this implies that such a life becomes inescapably transparent. If we consider this, our discussion about the privacy of photographs and wall posts barely scratches the surface. So what should we be discussing about? Probably, the three concepts that entitle this last section provide a good starting point.

Let us start with privacy. In 2010, Mark Zuckerberg, founder of Facebook, has claimed that privacy "is no longer a social norm" [189]. He proposes that through the use of the Internet and applications like Facebook, people (or at least some of them) have changed their perspectives on privacy, being more open to the idea of sharing information that would previously have been considered strictly private. Zuckerberg's proposals were not only well-received, and he was criticized for propagating a new conception of privacy to only benefit the business of his company, which could be described as profiting from a decreased privacy awareness. Regardless of Zuckerberg's personal involvement, the author believes that his thesis about our conception of privacy changing has some truth to it. The Internet has become a place where billions of interactions between people take place, and although many parts of it are protected from access by the general public, it can still be considered a rather public space. To return to the metaphor of the global village, we could say that although you can meet other villagers in their private homes, you still have to walk through public streets to get to them, and eventually, some of what is discussed in back rooms will end up on the village square. So on a factual level, the level of publicity of (some of) one's private matters has increased. And this in turn is apt to change our conception of privacy: once private matters are public, we may either try to make them non-public again, or simply accept that they are public and change our preferences. It would be wrong, however, to follow from this that individuals have no right to their individual conception of privacy. Based on the example of some individuals having a clearly low level of privacy concerns, one cannot conclude that this applies to our society as a whole. It should be possible for everyone to have their own understanding of privacy, and to enforce it against opposing forces. This brings us to the next point: control.

If one has a personal computer and stores data about oneself on it, one can make full use of the data, one can control who has access to it, and one can delete the data. But as has been stated before, data about oneself is no longer limited to the personal computer. First, some of the data one creates is stored at a location where one has only limited control over it. This applies to email services such as Gmail, shared document services, photo sharing sites, and a multitude of online applications that allow the creation of content through a web browser. Regarding control of this data, the questions of who

(or what) else is using the data, and of who can (definitively) delete or lock the data are highly relevant. The same questions arise in a case where the data about oneself does not result from the user creating herself, but rather from some form of external monitoring. Examples are credit and consumer fidelity card data, internet service provider traffic data or data mobile phone carriers acquire from their user's cellphones. In both cases, the physical control of the data is out of the user's hands, and in the second case, the user is not even the author of the data, and thus does not have the same rights as with self-authored data. Of course, in any of these cases, the user agrees on terms and conditions that govern the manner in which the data gathered by or about him are treated. Should a user in principal object to such terms, he should not use the service. That the finding of suitable terms is not trivial, and that a change in terms may have a significant impact, is reflected by a change in privacy terms on the part of Facebook. In 2009, with a count of 350 million users, Facebook published new privacy setting controls, in part as a reaction to criticism of its previous privacy settings [190]. The reactions were not positive, and one main point of criticism was that a user's friend list – i.e. a list of people who the user is friends with – was now considered as *public information*. Previously, users had been able to only share their friend list with their friends. In the new version, anybody on the Internet was able to see who someone is friends with. Facebook has continued to evolve its privacy settings, and much of the criticism has been addressed. This example should serve to show two things. First, the interests of the actors involved in cases where data is not in full control of the person it is about may have diverging interests. Second, and more importantly, the modalities of control are by no means set, and they are subject to ongoing refinement. This means that as social media, and more generally speaking, electronic memory continue to gain more relevance in our lives, changes to the manner in which privacy settings are controlled will go along with that development. That these changes will be significant is shown by the example of suddenly considering a friend list as a public asset. Whom one is friends with, and whom not, should not be misunderstood as a demonstration of how popular one is. Rather, it provides a deep insight into one's private life.

So far, the actors in our discussion were individuals about whom the data is, and service providers that are somehow involved in the creation of this data. We can see the problems discussed so far as exemplified in every relationship between an individual and such a service provider. What we have not seen so far, however, is the integration of all these relationships. If such an integration were possible, the consequences with respect to an individual's privacy could, in case there was misuse, be the most severe. We consider that the only actor that potentially able to perform such an integration is the state. It is at this point that the concept of the transparent citizen comes into play.

The term *surveillance*, from an individual's perspective, refers to someone else – be it a person or an institution – collecting data about the individual. A good example of surveillance, its utility and its criticism is public video

surveillance. If the state places video cameras in public places and records the events that take place, it may later use these videos as evidence and to identify alleged delinquents. While the state may perceive such systems as a contribution to public safety, individuals may perceive them as a threat to their privacy and freedom. In any way, the coverage of an individual's life by such systems will always remain very limited. This is not the case, however, for electronic memory. In the eyes of some of the proponents of lifelogging and electronic memory, the total capture of an individual's life in the form of digital information is desirable [7]. At first sight, this seems to make sense: if we want to make full use of all the digital information we acquire over a lifetime, it should be as complete and encompassing as possible. From the perspective of someone performing surveillance, such a complete record of an individual is more than he could ever ask for. It is for this reason that lifelogging and the acquisition of electronic memory is sometimes termed as *sousveillance*. The term was coined by Mann [191] to describe the activity of surveying the surveyors, but has more recently been used to refer to the potential threat one's electronic memory can pose to one self. Allen has stated that there is no reason to believe that states will restrain themselves from trying to access individual's electronic memory under certain circumstances, and has called the *sousveillor* "the true sibling of Big Brother" ([192], p. 20). Allen notes that in the United States, personal diaries have been admitted as evidence against their authors in court, and she sees no reason why the case should be different for electronic memory.

While electronic memory has been introduced as a convenient service of a future computing infrastructure, capable of improving our lives, there is a danger that the use of such technologies will make us into something we could call *transparent citizens*. Such citizens would be citizens for which the state or any other actor with comparable power could obtain a nearly complete digital record of their lives, being free to analyze and evaluate it with whatever goal in mind. Such a scenario is frightening, but it should not be seen as inevitable. The development of electronic memory is ongoing, and the future infrastructure it will bring about will find a way to make electronic memory beneficial to its users, and not a potential threat. On a conceptual level, the solution is astonishingly simple. The functions of our biological memory modeled so far all focus on retrieving information – reflection, recollection, reminiscence, retrieval and remembering intentions. All we must do is include what should, given the previous scenario, probably be seen as the most important function of our memory: forgetting.







## References

- [1] M. Weiser. The computer for the twenty-first century. *Scientific American*, 265(3):94–104, 1991.
- [2] V. Bush. As We May Think. *The Atlantic Monthly*, 176(1):101–108, 1945.
- [3] M.K. Buckland. Emanuel Goldberg, electronic document retrieval, and Vannevar Bush's Memex. *Journal of the American Society for Information Science*, 43(4):284–294, 1992.
- [4] S. Mann. Wearable computing: A first step toward personal imaging. *Computer*, 30(2):25–32, 1997.
- [5] K. Achilleos. Evolution of Lifelogging. [http://mms.ecs.soton.ac.uk/2010/papers/Evolution\\_of\\_Lifelogging.pdf](http://mms.ecs.soton.ac.uk/2010/papers/Evolution_of_Lifelogging.pdf), accessed October 2011, 2008.
- [6] J. Gemmell, G. Bell, R. Lueder, S. Drucker, and C. Wong. MyLifeBits: fulfilling the Memex vision. In *Proceedings of the tenth ACM international conference on Multimedia*, page 238. ACM, 2002.
- [7] J. Gemmell, G. Bell, and R. Lueder. MyLifeBits: a personal database for everything. *Communications of the ACM*, 49(1):95, 2006.
- [8] A.J. Sellen, A. Fogg, M. Aitken, S. Hodges, C. Rother, and K. Wood. Do life-logging technologies support memory for the past?: an experimental study using sensecam. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, page 90. ACM, 2007.
- [9] A.J. Sellen and S. Whittaker. Beyond total capture: a constructive critique of lifelogging. *Communications of the ACM*, 53(5):70–77, 2010.
- [10] N. Stephenson. *Snow crash*. ePenguin, 1992.
- [11] Linden Labs. Second Life. <http://secondlife.com>, accessed October 2011, 2011.
- [12] Virtual Business Management Ltd. European Linden Dollar Exchange. <https://www.eldexchange.eu>, accessed October 2011, 2011.
- [13] R. Hof. Second Life's First Millionaire. [http://www.businessweek.com/the\\_thread/techbeat/archives/2006/11/second\\_lifes\\_fi.html](http://www.businessweek.com/the_thread/techbeat/archives/2006/11/second_lifes_fi.html), accessed October 2011.
- [14] R. Harper. Clarification. <http://couldtherebewhales.blogspot.com/2009/03/correction.html>, accessed October 2011, 2009.

- [15] Stanford University. Metaverse U Conference. <http://metaverse.stanford.edu/>, accessed October 2011, 2009.
- [16] P. O'Mahony. Sweden trumped by Maldives in Second Life. 2007.
- [17] EJ Smart, J. Cascio, and J. Paffendorf. Metaverse Roadmap Overview. <http://www.metaverseroadmap.org>, accessed October 2011, 2007.
- [18] The Internet Archive. The Internet Archive: Universal Access to all knowledge. <http://www.archive.org>, accessed October 2011.
- [19] R. Want and T. Pering. New horizons for mobile computing. In *Pervasive Computing and Communications, 2003.(PerCom 2003). Proceedings of the First IEEE International Conference on*, pages 3–8. IEEE.
- [20] F. Müller, P. Fornaro, L. Rosenthaler, and R. Gschwind. Peviar: Digital originals. *Journal on Computing and Cultural Heritage (JOCCH)*, 3(1):2, 2010.
- [21] A. Amir, F. Müller, P. Fornaro, R. Gschwind, J. Rosenthal, and L. Rosenthaler. Towards a channel model for microfilm. *Archiving 2008*, 5:207–211, 2008.
- [22] F. Müller, M. Guggisberg, and H. Burkhart. Email as Electronic Memory: A Spatial Exploration Interface. In *The Third International Conferences on Advances in Multimedia*, pages 98–103, 2011.
- [23] Jan Lewe Torpus. Lifeclipper2 Project Page. <http://www.lifeclipper2.idk.ch/>, accessed October 2011, 2011.
- [24] F. Müller. Lifeclipper3: Massively Augmented Reality and its Realization using Community Processes. In *InterMedia Summer School*, 2009.
- [25] T. Denzler. HUVis: Handheld Urban Visualization (Master Thesis), 2011.
- [26] Reagan Moore. Towards a Theory of Digital Preservation. *The International Journal of Digital Curation*, 3(1):63–75, 2008.
- [27] M. Mois, C.P. Klas, and M. Hemmje. Digital preservation as communication with the future. In *DSP 2009, Proceedings of the Fifth International Conference on Web Information Systems and Technologies*, 2009.
- [28] CCSDS. OAIS (Reference Model for an Open Archival Information System), 2002. Consultative Committee for Space Data Systems.
- [29] ISO. ISO 14721:2003. Space data and information transfer systems – Open archival information system – Reference model, 2003.
- [30] J. Rothenberg. Ensuring the longevity of digital documents. *Scientific American*, 272(1):42–47, 1995.

- [31] T. Kuny. The Digital Dark Ages? Challenges in the Preservation of Electronic Information. *International Preservation News*, pages 8–13, 1998.
- [32] M. Hedstrom. Digital preservation: a time bomb for digital libraries. *Computers and the Humanities*, 31(3):189–202, 1997.
- [33] O. Slattery, R. Lu, J. Zheng, F. Byers, and X. Tang. Stability Comparison of Recordable Optical Discs—A Study of Error Rates in Harsh Conditions. *Journal of Research - National Institute of Standards and Technology*, 109(5):517, 2004.
- [34] Daniela Bienz, Rudolf Gschwind, Mario Pozzaand, and Ludwig Gantner. Mass CD/DVD Migration: A Novartis Case Study. In *IS&T's Archiving Conference Proceedings*, 2010.
- [35] Michele Youket and Nels Olson. Compact Disc Service Life Studies by the Library of Congress. In *IS&T's Archiving Conference Proceedings*, pages 99–104, 2007.
- [36] Brack. Brack Online Store. <http://www.brack.ch>, accessed October 2011, 2011.
- [37] Nick Zaoino Douglas Stinson, Fred Ameli. Lifetime of KODAK Writable CD and Photo CD Media. <http://www.cd-info.com/CDIC/Technology/CD-R/Media/Kodak.html>, accessed October 2011, 1995.
- [38] Fred Byers. Care and Handling of CDs and DVDs – A Guide for Librarians and Archivists. *NIST Special Publication 500-252*, 2003.
- [39] Chandru Shanani Basil Manns. Longevity of CD Media Research at the Library of Congress. <http://www.loc.gov/preserv/sutdyofCDlongevity.pdf>, accessed October 2011, 2003.
- [40] Stefan Rohde-Enslin. Nicht von Dauer. Kleiner Ratgeber für die Bewahrung digitaler Daten in Museen. <http://www.langzeitarchivierung.de>, accessed October 2011, 2004.
- [41] Kevin Bradley. Defining Digital Sustainability. *Library Trends*, 56(1):148–163, 2007.
- [42] D.S.H. Rosenthal. Bit Preservation: A Solved Problem? In *Proceedings of iPres*. Stanford University, 2008.
- [43] G.E. Moore. Cramming more components onto integrated circuits, *Electronics*. *April*, 19:114–117, 1965.
- [44] Fujifilm. DLT Media and Format – Drive Compatibility Guide. <http://lib.store.yahoo.net/lib/tapestock/DLTFormatCompatibilityGuide.pdf>, accessed October 2011, 2006.

- [45] Sergio Gregorio. Defining Digital Archeology. In *IS&T's Archiving Conference Proceedings*, pages 92–95, 2009.
- [46] Christ Rusbridge. Excuse Me...Some Digital Preservation Fallacies? *Ariadne*, 2006.
- [47] Lukas Rosenthaler. *Archivierung im digitalen Zeitalter. Historische Entwicklung und Wege in eine digitale Zukunft*. University of Basel, 2006.
- [48] Jeff Rothenberg. *Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation. A Report to the Council on Library and Information Resources*. Council on Library and Information Resources, 1999.
- [49] U.M. Borghoff, P. Rödiger, J. Scheffczyk, and L. Schmitz. *Langzeitarchivierung: Methoden zur Erhaltung digitaler Dokumente*. Dpunkt. Verlag, 2003.
- [50] Susanne Dobratz. Grundfragen der digitalen Langzeitarchivierung für den edoc-Server. *CMD-Journal*, 2009.
- [51] Library of Congress. The Deterioration and Preservation of Paper: Some Essential Facts. <http://www.loc.gov/preserv/deterioratebrochure.html>, accessed October 2011, 2010.
- [52] Urs Leu. Handbuch der historischen Buchbestände in der Schweiz. <http://hhch.europider.com/spezielsammlungen/alte-drucke-rara/handbuchhistorisch/index.html>, accessed October 2011, 2010.
- [53] Norman Woodland, N.J. Ventnor, and Bernard Silver. Classifying Apparatus and Method. U.S. Patent 2.612.994, 1952.
- [54] ISO. Information technology – Automatic identification and data capture techniques – QR Code 2005 bar code symbology specification, 2006.
- [55] J.H. Altman and H.J. Zweig. Effect of Spread Function on the Storage of Information on Photographic Emulsions. *Photographic Science and Engineering*, 7(3):173–177, 1963.
- [56] J. Cheney, C. Lagoze, and P. Botticelli. Towards a Theory of Information Preservation. *Research and Advanced Technology for Digital Libraries*, pages 340–351, 2001.
- [57] Y. Tzitzikas and G. Flouris. Mind the (Intelligibility) Gap. *LECTURE NOTES IN COMPUTER SCIENCE*, 4675:87, 2007.
- [58] *The Digital Time Capsule and Other Applications of Microfilm*, 2010.

- [59] Joel Daavid. *The History of Microfilm*, 2010.
- [60] Ilford AG. Ilfochrome Micrographic. [http://www.ilford.com/en/pdf/prods/micrographic/MICRO\\_graphic\\_chemistry.pdf](http://www.ilford.com/en/pdf/prods/micrographic/MICRO_graphic_chemistry.pdf), accessed October 2011, 2002.
- [61] ECMA International. Test Method for the Estimation of the Archival Lifetime of Optical Media. <http://www.ecma-international.org/publications/standards/Ecma-379.htm>, accessed October 2011, 2008.
- [62] A. Meyer and D. Bermane. Silver Dye-Bleach Color Microfilm. *Journal of Applied Photographic Engineering*, 9(4):117–120, 1983.
- [63] A. Meyer and D. Bermane. The stability and permanence of cibachrome images. *Journal of Applied Photographic Engineering*, 9(4):117–120, 1983.
- [64] H.G. Wilhelm and C. Brower. *The permanence and care of color photographs: traditional and digital color prints, color negatives, slides, and motion pictures*. Preservation Publishing Company, Grinnell (US-IA), 1993.
- [65] E.W. Pugh, L.R. Johnson, and J.H. Palmer. *IBM's 360 and Early 370 Systems*. The MIT Press, 1991.
- [66] Dolby. Dolby Digital Sound. <http://www.dolby.com>, accessed October 2011, 2010.
- [67] Sony. SDDS Laboratory Process Manual V3.2. <http://www.sdds.com/pdfs/SDDSLabQuickGuide.pdf>, accessed October 2011, 2004.
- [68] T. Van der Veen and J.M. Clifton. Method, Apparatus and Software for Processing Photographic Image Data Using a Photographic Recording Medium. US Patent Application Publication no. US 2008/0292300 A1, 2008.
- [69] D. Gubler, L. Rosenthaler, and P. Fornaro. The Obsolescence of Migration: Long-Term Storage of Digital Code on Stable Optical Media. In *IS&T's Archiving Conference Proceedings*, pages 135–139. IS&T, 2006.
- [70] A. Hofmann and D. Gield. Long Term Migration Free Storage of Digital Audio Data on Microfilm. In *IS&T's Archiving Conference Proceedings*, pages 184–187. IS&T, 2008.
- [71] C. Voges and T. Fingscheidt. Technology and Application of Digital Data Storage on Microfilm. *Journal of Imaging Science and Technology*, 53:060505–1 – 060505–8, 2009.

- [72] Claude E. Shannon. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27:379–423, 625–56, 1948.
- [73] P. Glafkidès. *Chimie et physique photographiques*. Éditions de l'Usine Nouvelle, 1987.
- [74] R. Shaw. The Application of Fourier Techniques and Information Theory to the Assessment of Photographic Image Quality. *Photographic Science and Engineering*, 6(5):281–286, 1962.
- [75] A. Amir, F. Müller, P. Fornaro, R. Gschwind, J. Rosenthal, and L. Rosenthaler. Towards a Channel Model for Microfilm. In *IS&T's Archiving Conference Proceedings*, pages 207–211, 2008.
- [76] ISO. ISO 12233:2000(E): Photography - Electronic still-picture cameras - Resolution measurements, 2000.
- [77] Peter Burns. sfrmat2. <http://losburns.com/imaging/software/SFRedge/index.htm>, accessed October 2011, 2003.
- [78] EWH Selwyn. A theory of graininess. *Photog. J*, 73:571, 1935.
- [79] Ariel Amir. *PhD Thesis (forthcoming)*. PhD thesis, Institute for Mathematics, University of Zurich, 2010.
- [80] Jay Jacobsmeyer. Introduction to Error-Control Coding. [http://www.pericle.com/papers/Error\\_Control\\_Tutorial.pdf](http://www.pericle.com/papers/Error_Control_Tutorial.pdf), accessed October 2011, 2004.
- [81] B.A. Cipra. The ubiquitous reed-solomon codes. *SIAM News*, 26(1), 1993.
- [82] Peter Rechenberg and Gustav Pomberger. *Informatik-Handbuch*. Hanser, 1999.
- [83] Andrew Wilson. Letter to the Editor: Authentic Digital Objects. *The International Journal of Digital Curation*, 4, 2009.
- [84] C.T. Cullen. Authentication of digital objects: Lessons from a Historian's Research. *Authenticity in a digital environment*, pages 1–7, 2000.
- [85] D.M. Levy. Wheres Waldo? Reflections on Copies and Authenticity in a Digital Environment. *Authenticity in a Digital Environment*, pages 24–31, 2000.
- [86] H. Neuroth, A. Osswald, and others (Ed.). *nestor Handbuch. Eine kleine Enzyklopädie der digitalen Langzeitarchivierung*. Verlag Werner Huelsbusch, Boizenburg, 2009.

- [87] ISO. ISO 15489:2001-1/2. Information and documentation – Records management – Part 1: General, Part 2: Guidelines, 2001.
- [88] J. Rothenberg. Preserving authentic digital information. *Authenticity in a digital environment*, pages 51–68, 2000.
- [89] Filip Boudrez. Digital Signatures and Electronic Records. *Archival Science*, 7(2):179–193, 2007.
- [90] M. Stevens, A. Sotirov, J. Appelbaum, A. Lenstra, D. Molnar, D.A. Osvik, and B. de Weger. Short chosen-prefix collisions for md5 and the creation of a rogue ca certificate. In *Proceedings of the 29th Annual International Cryptology Conference on Advances in Cryptology*, page 69. Springer, 2009.
- [91] C. Lynch. Authenticity and integrity in the digital environment: An exploratory analysis of the central role of trust. *Authenticity in a digital environment*, pages 32–50, 2000.
- [92] F.Y. Wang, K.M. Carley, D. Zeng, and W. Mao. Social computing: From social informatics to social intelligence. *Intelligent Systems, IEEE*, 22(2):79–83, 2007.
- [93] Wikimedia Foundation. Wikipedia – The Online Encyclopedia. <http://www.wikipedia.org>, accessed October 2011, 2011.
- [94] Inc. Facebook. Facebook. <http://www.facebook.com>, accessed October 2011, 2011.
- [95] Leena Rao. Twitter Seeing 90 Million Tweets Per Day, 25 Percent Containing Links. *TechCrunch*, 2010.
- [96] Inc. Google. DoubleClick Ad Planner by Google. <http://www.google.com/adplanner>, accessed October 2011, 2011.
- [97] Nick. Social Media Demographics. <http://avvoblog.com/2010/04/29/social-media-demographics/>, accessed October 2011, 2010.
- [98] J. Riedl. The Promise and Peril of Social Computing. *IEEE Computer*, 44(1):93–96, 2011.
- [99] I. Safrin and C.M. Schmidt. Pastiche. <http://www.christianmarcschmidt.com/projects/pastiche/>, accessed October 2011, 2009.
- [100] S. Kamvar and J. Harris. We Feel Fine and Searching the Emotional Web. In *4th ACM International Conference on Web Search and Data Mining*, pages 117–126. ACM, 2011.

- [101] John Scott. *Social Network Analysis - A Handbook*. SAGE Publications, 2000.
- [102] Bruno Preiss. *Data Structures and Algorithms with Object-Oriented Design Patterns in C++*. Albazaar, 1999.
- [103] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang. Complex networks: Structure and dynamics. *Physics Reports*, pages 175–308, 2006.
- [104] Linton C. Freeman. *The Development of Social Network Analysis*. Empirical Press, 2004.
- [105] S. Milgram. The Small World Problem. *Psychology Today*, 2(1):60–67, 1967.
- [106] W.W. Zachary. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, pages 452–473, 1977.
- [107] Anonymous. Stanford Large Network Dataset Collection. <http://snap.stanford.edu/>, accessed October 2011, 2011.
- [108] List of virtual communities with more than 100 million users. <http://en.wikipedia.org/wiki/etc.>, accessed October 2011, 2011.
- [109] Inc. Facebook. Facebook Statistics. <http://http://www.facebook.com/press/info.php?statistics>, accessed October 2011, 2011.
- [110] S. Eubank, H. Guclu, V.S.A. Kumar, M.V. Marathe, A. Srinivasan, Z. Toroczkai, and N. Wang. Modelling disease outbreaks in realistic urban social networks. *Nature*, 429(6988):180–184, 2004.
- [111] T.W. Valente. Social network thresholds in the diffusion of innovations. *Social Networks*, 18(1):69–89, 1996.
- [112] J.H. Fowler and N.A. Christakis. Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the Framingham Heart Study. *British Medical Journal*, 337(2338):1–9, 2008.
- [113] L.F. Berkman and S.L. Syme. Social networks, host resistance, and mortality: a nine-year follow-up study of Alameda County residents. *American journal of Epidemiology*, 109(2):186, 1979.
- [114] S. Ressler. Social network analysis as an approach to combat terrorism: past, present, and future research. *Homeland Security Affairs*, 2(2):1–10, 2006.
- [115] D. Hansen, B. Shneiderman, and M. Smith. Visualizing Threaded Conversation Networks: Mining Message Boards and Email Lists for Actionable Insights. *Active Media Technology*, pages 47–62, 2010.



- [116] Federal Regulatory Energy Commission. Information Released in Enron Investigation. <http://www.ferc.gov/industries/electric/industry/wec/enron/info-release.asp>, 2010.
- [117] John Wang. The Enron Data Reconstruction Project. <http://enrondata.org>, accessed October 2011, 2011.
- [118] B. Keller. Enron for Dummies. <http://www.nytimes.com/2002/01/26/opinion/enron-for-dummies.html>, accessed October 2011, 2002.
- [119] Gilardi Co. Llc. Enron Historical Share Price Table. <http://www.gilardi.com/enron/securities/>, accessed October 2011, 2011.
- [120] Anonymous. Wall Street Journal Enron Glossary. <http://online.wsj.com/public/resources/documents/info-enrongloss-0603.html>, accessed October 2011, 2006.
- [121] D.K. Berman. Online Laundry: Government Posts Enron's E-Mail. *The Wall Street Journal*, 2003.
- [122] K. Krasnow Waterman. Knowledge Discovery in Corporate Email: The Compliance Bot Meets Enron. Master's thesis, MIT, 2006.
- [123] W.W. Cohen. Enron Email Dataset. <http://www.cs.cmu.edu/~enron/>, accessed October 2011, 2009.
- [124] Anonymous. EDRM Enron PST Data Set. <http://edrm.net/resources/data-sets/enron-data-set-files>, accessed October 2011, 2011.
- [125] A. McCallum, X. Wang, and A. Corrada-Emmanuel. Topic and role discovery in social networks with experiments on enron and academic email. *Journal of Artificial Intelligence Research*, 30(1):249–272, 2007.
- [126] J. Shetty and J. Adibi. The Enron email dataset database schema and brief statistical report. *Information Sciences Institute Technical Report, University of Southern California*, 2004.
- [127] J. Diesner, T.L. Frantz, and K.M. Carley. Communication Networks from the Enron Email Corpus It's Always About the People. Enron is no Different. *Computational & Mathematical Organization Theory*, 11(3):201–228, 2005.
- [128] Y. Zhou, M. Goldberg, M. Magdon-Ismail, and W.A. Wallace. Strategies for cleaning organizational emails with an application to enron email dataset. In *5th Conf. of North American Association for Computational Social and Organizational Science*. Citeseer, 2007.

- [129] M.W. Berry, M. Browne, and B. Signer. 2001 Topic Annotated Enron Email Data Set. LDC Catalog No. LDC2007T22, 2007.
- [130] R. Bekkerman, A. McCallum, and G. Huang. Automatic Categorization of Email into Folders: Benchmark Experiments of Enron and SRI Corpora. Technical Report IR-418, University of Massachusetts, 2004.
- [131] B. Klimt and Y. Yang. The Enron Corpus: A New Dataset for Email Classification Research. *Machine Learning: ECML 2004*, pages 217–226, 2004.
- [132] R. Nussbaum, A.H. Esfahanian, and P.N. Tan. History-Based Email Prioritization. In *Social Network Analysis and Mining, 2009. ASONAM'09. International Conference on Advances in*, pages 364–365. IEEE, 2009.
- [133] V.R. Carvalho and W. Cohen. Recommending recipients in the Enron email corpus. 2007.
- [134] T. Ayodele, S. Zhou, and R. Khusainov. Email Reply Prediction: A Machine Learning Approach. *Human Interface and the Management of Information*, pages 114–123, 2009.
- [135] G. Creamer, R. Rowe, S. Hershkop, and S. Stolfo. Segmentation and automated social hierarchy detection through email network analysis. *Advances in Web Mining and Web Usage Analysis*, pages 40–58, 2009.
- [136] A. McCallum, X. Wang, and A. Corrada-Emmanuel. Topic and role discovery in social networks with experiments on enron and academic email. *Journal of Artificial Intelligence Research*, 30(1):249–272, 2007.
- [137] L. Hossain and A. Wu. Communications network centrality correlates to organisational coordination. *International Journal of Project Management*, 27(8):795–811, 2009.
- [138] G. Di Battista, P. Eades, R. Tamassia, and I.G. Tollis. *Graph Drawing: Algorithms for the Visualization of Graphs*. Prentice Hall PTR Upper Saddle River, NJ, USA, 1998.
- [139] T.M.J. Fruchterman and E.M. Reingold. Graph drawing by force-directed placement. *Software – Practice and Experience*, 21:1129–1164, 1991.
- [140] A. Frick, A. Ludwig, and H. Mehldau. A fast adaptive layout algorithm for undirected graphs (extended abstract and system demonstration). In *Graph Drawing*, pages 388–403. Springer, 1995.
- [141] Markus M. Geipel. *Dynamics of Communities of Code in Open Source Software*. PhD thesis, ETH Zurich, 2010.

- [142] J.T. Bernstein. Traer Physics 3.0. <http://murderandcreate.com/physics/>, 2009.
- [143] J. Shetty and J. Adibi. Ex Employee Status Report. [http://www.isi.edu/~adibi/Enron/Enron\\_Employee\\_Status.xls](http://www.isi.edu/~adibi/Enron/Enron_Employee_Status.xls), accessed October 2011, 2003.
- [144] Mark Maney. Employee Comittee Complaint against John D. Arnold et al. [http://archive.employeecommittee.org/pdf/56\\_defendants.pdf](http://archive.employeecommittee.org/pdf/56_defendants.pdf), accessed October 2011, 2003.
- [145] Robert McCullough. Memorandum – Reading Enron’s Scheme Account-ing Materials. [www.mresaerch.com](http://www.mresaerch.com), accessed October 2011, 2004.
- [146] Fernanda Viégas. Visualizations of PostHistory. <http://alumni.media.mit.edu/~fviegas/posthistory/vis.html>, accessed October 2011, 2002.
- [147] Simone Frau, Jonathan Robert, and Nadia Boukhelifa. Dynamic Coor-dinated Email Visualization. In *WSCG’05 Proceedings*, 2005.
- [148] F.B. Viégas, S. Golder, and J. Donath. Visualizing email content: por-traying relationships from conversational histories. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 979–988. ACM, 2006.
- [149] F.B. Viégas, D. Boyd, D.H. Nguyen, J. Potter, and J. Donath. Digital artifacts for remembering and storytelling: posthistory and social network fragments. In *System Sciences, 2004. Proceedings of the 37th Annual Hawaii International Conference on*, page 10. IEEE, 2004.
- [150] E. Freeman and D. Gelernter. Lifestreams: a storage model for personal data. *ACM SIGMOD Record*, 25(1):80–86, 1996.
- [151] J. Heer. Exploring Enron – Visual Data Mining of E-Mail. <http://hci.stanford.edu/jheer/projects/enron/>, accessed October 2011, 2005.
- [152] R. Descartes and D.A. Cress. *Discourse on method; Meditations on first philosophy*. Hackett Publishing Company, 1998.
- [153] A. Wachowski and L. Wachowski. *The Matrix*, 1999.
- [154] D. Cronenberg. *eXistenZ*, 1999.
- [155] I.E. Sutherland. The ultimate display. In *Proceedings of the IFIP Congress*, volume 2, pages 506–508. Citeseer, 1965.

- [156] P. Milgram, H. Takemura, A. Utsumi, and F. Kishino. Augmented Reality: A Class of Displays on the Reality-Virtuality Continuum. In *Proceedings of Telemanipulator and Telepresence Technologies*, volume 2351, pages 282–292. Citeseer, 1994.
- [157] R.T. Azuma et al. A Survey of Augmented Reality. *Presence-Teleoperators and Virtual Environments*, 6(4):355–385, 1997.
- [158] J.P. Rolland and H. Fuchs. Optical versus video see-through head-mounted displays in medical visualization. *Presence: Teleoperators & Virtual Environments*, 9(3):287–309, 2000.
- [159] Imagination Computer Services. iotracker. <http://www.iotracker.com>, accessed October 2011, 2011.
- [160] H. Kato. ARToolkit. <http://www.hitl.washington.edu/artoolkit/>, accessed October 2011, 2001.
- [161] Studierstube Tracker. <http://handheldar.icg.tugraz.at/stbtracker.php>, accessed October 2011, 2011.
- [162] Inc. AR Toolworks. Artoolkit nft. <http://www.artoolworks.com/products/stand-alone/artoolkit-nft/>, accessed October 2011, 2011.
- [163] G. Papagiannakis, G. Singh, and N. Magnenat-Thalmann. A survey of mobile and wireless technologies for augmented reality systems. *Computer Animation and Virtual Worlds*, 19(1):3–22, 2008.
- [164] W. Piekarski and B.H. Thomas. The tinmith system: demonstrating new techniques for mobile augmented reality modelling. In *Proceedings of the Third Australasian conference on User interfaces-Volume 7*, pages 61–70. Australian Computer Society, Inc., 2002.
- [165] A.D. Cheok, K.H. Goh, W. Liu, F. Farbiz, S.W. Fong, S.L. Teo, Y. Li, and X. Yang. Human Pacman: a mobile, wide-area entertainment system based on physical, social, and ubiquitous computing. *Personal and Ubiquitous Computing*, 8(2):71–81, 2004.
- [166] D. Wagner, T. Pintaric, F. Ledermann, and D. Schmalstieg. Towards massively multi-user augmented reality on handheld devices. *Pervasive Computing*, pages 208–219, 2005.
- [167] Wikitude. <http://www.wikitude.com>, accessed October 2011, 2011.
- [168] Layar. <http://www.layar.com>, accessed October 2011, 2011.
- [169] Bruce Sterling. At the Dawn of the Augmented Reality Industry. <http://vimeo.com/6189763>, accessed October 2011, 2009.

- [170] Leica Geosystems. Leica GPS1200 and RX1250 RCU. <http://www.leica-geosystems.com>, accessed October 2011, 2008.
- [171] Inertia Cube 3. <http://www.isense.com>, accessed October 2011, 2008.
- [172] Virtools. <http://www.virttools.com>, accessed October 2011, 2008.
- [173] Russel M. Taylor. VRPN - The Virtual Reality Peripheral Network. <http://www.cs.unc.edu/Research/vrpn>, accessed October 2011, 2008.
- [174] S. You, U. Neumann, and R. Azuma. Hybrid inertial and vision tracking for augmented reality registration. In *Virtual Reality, 1999. Proceedings., IEEE*, pages 260–267. IEEE, 1999.
- [175] O. Koch. Position-Dependent Filtering of Objects in an Augmented Reality Environment. <http://www.lifeclipper2.idk.ch/DE/publications.html>, accessed October 2011, 2009.
- [176] J. Torpus. Lifeclipper3. Script. <http://www.lifeclipper3.torpus.com/>, accessed October 2011, 2010.
- [177] D.G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2, pages 1150–1157. IEEE, 1999.
- [178] E.N. Mortensen, H. Deng, and L. Shapiro. A sift descriptor with global context. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 184–190. IEEE, 2005.
- [179] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1615–1630, 2005.
- [180] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.
- [181] D.M. Mount and S. Arya. ANN: A library for approximate nearest neighbor searching. In *CGC 2nd Annual Fall Workshop on Computational Geometry*, 1997.
- [182] POV-Ray - Persistence of Vision Raytracer. <http://www.povray.org>, accessed October 2011, 2011.
- [183] Y. Furukawa and J. Ponce. Accurate, Dense, and Robust Multi-View Stereopsis. In *PAMI 2010*, 2010.
- [184] M. Lösler. Java Graticule 3D. <http://derletztekick.com/software/netzausgleichung>, accessed October 2011, 2011.

- [185] P. Debevec. Rendering synthetic objects into real scenes: Bridging traditional and image-based graphics with global illumination and high dynamic range photography. In *ACM SIGGRAPH 2008 classes*, pages 1–10. ACM, 2008.
- [186] Facebook 8 developer conference. <http://www.facebook.com/f8>, accessed October 2011, 2011.
- [187] EU. Directive 2006/24/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL. *Official Journal of the European Union*, 105(54):2006, 2006.
- [188] K. Biermann. Was Vorratsdaten über uns verraten. *ZEIT Online*, 2011.
- [189] E. Barnett. Facebook’s Mark Zuckerberg says privacy is no longer a ‘social norm’. *The Telegraph*.
- [190] K. Bankston. Facebook’s New Privacy Changes: The Good, The Bad, and The Ugly. <https://www.eff.org/deeplinks/2009/12/facebooks-new-privacy-changes-good-bad-and-ugly>, accessed October 2011, 2009.
- [191] S. Mann, J. Nolan, and B. Wellman. Sousveillance: Inventing and using wearable computing devices for data collection in surveillance environments. *Surveillance and society*, 1(3):331–355, 2003.
- [192] A.L. Allen. Dredging up the past: Lifelogging, memory, and surveillance. *The University of Chicago Law Review*, pages 47–74, 2008.
- [193] Hellmut Frieser. *Photographische Informationsaufzeichnung*. R. Oldenbourg and Focal Press, 1975.

## List of Figures

1	The metaverse continuum space . . . . .	19
2	Digital Preservation Communication model . . . . .	30
3	The problematic triangle of digital archiving . . . . .	33
5	Level of concern of carrier decay . . . . .	34
4	Analogue and digital information degradation . . . . .	35
6	Assortment of Storage Technologies . . . . .	40
7	Quantum DLT Compatibility Chart . . . . .	41
8	Level of concern of hardware obsolescence . . . . .	42
9	Level of concern of software obsolescence . . . . .	44
10	Comparison of one- and two-dimensional barcodes . . . . .	55
11	VGA fonts . . . . .	55
12	Illustration of Self-documentation . . . . .	56
13	The Pioneer plaque . . . . .	57
14	General Communication System . . . . .	64
15	The effect of granularity on sharpness . . . . .	65
16	Limited modulation transfer . . . . .	66
17	Peviar microfilm example . . . . .	69
18	Illustration of the Peviar workflow . . . . .	70
19	Microfilm as a general communication system . . . . .	70
20	Edge sharpness: Razor blade . . . . .	72
21	Edge sharpness: Vacuum-deposited edge . . . . .	72
22	Edge sharpness: Laser-recorded edge . . . . .	73
23	Modulation transfer function for Peviar . . . . .	74
24	Peviar MTF tabular report . . . . .	75
25	Relative RMS granularity of color microfilm . . . . .	77
26	Magenta dye layer at various exposures . . . . .	78
27	Error correction code workflow . . . . .	81
28	Individual barcode cell . . . . .	83
29	Peviar write process . . . . .	85
30	Peviar read process . . . . .	86
31	MD5 command line tool hash . . . . .	91
32	Asymmetric cryptographic protocol applications . . . . .	93
33	Twitter-based Social Memory I . . . . .	102
34	Twitter-based Social Memory II . . . . .	102
35	Simple unweighted, directed graph . . . . .	104
36	Simple weighted, directed graph . . . . .	104
37	Graph Density . . . . .	105
38	Network Centralization . . . . .	106
39	Different centrality measures . . . . .	107
40	Grouping in graphs . . . . .	109
41	Clustering Illustration . . . . .	110
42	Agglomerative Clustering . . . . .	111

43	Enron Historical Stock . . . . .	118
44	Enron Corpus Message Distribution . . . . .	123
45	Sent emails per custodian . . . . .	124
46	Distribution number of messages over hour of day . . . . .	125
47	Arbitrary graph drawings . . . . .	130
48	Illustration of network relaxation . . . . .	133
49	Illustration of the Reaction-Diffusion Model . . . . .	135
50	Discrete growth and continuous ageing . . . . .	136
51	Overall system architecture . . . . .	139
52	Enron MySQL database structure . . . . .	140
53	Database to business object mapping . . . . .	141
54	Enron simulation visualization 1 . . . . .	143
55	Enron simulation visualization 2 . . . . .	143
56	Two-step simulation illustration . . . . .	146
57	Agglomerative Clustering of the Enron graph . . . . .	148
58	Interactive Clustering Tool . . . . .	149
59	Clustering Tool . . . . .	150
60	Partial Organizational Chart 2000-52 . . . . .	151
61	Partial Organizational Chart 2001-52 . . . . .	152
62	PostHistory Email Interface . . . . .	153
63	Enron Social Graph Visualization I . . . . .	155
64	Enron Social Graph Visualization II . . . . .	156
65	Personal Email: Social Graph . . . . .	157
66	Personal Email: Information Forwarding . . . . .	158
67	Mixed Reality Illustration . . . . .	165
68	Mixed reality continuum . . . . .	166
69	Photograph of a head-mounted video see-through display . . . . .	168
70	Illustration of the tracking problem . . . . .	169
71	Lifeclipper2 System . . . . .	172
72	Lifeclipper2 Scenarios . . . . .	174
73	Lifeclipper2 Schematic . . . . .	175
74	Lifeclipper2 System Architecture . . . . .	176
75	Hybrid visualization of a future building . . . . .	179
76	HUVis Workflow . . . . .	181
77	Pointcloud 'grid' set input images . . . . .	182
78	Decision criteria for feature matching . . . . .	183
79	Illustration of point cloud by two example sets . . . . .	184
80	LSZ Model . . . . .	185
81	Rendering of building and occlusion environment . . . . .	186
82	Overlay of point cloud, model and photograph . . . . .	187
83	Hybrid image with omnidirectional panorama lighting . . . . .	188
84	Schematic overview of the metaverse archive . . . . .	198
85	The effect of the aperture on granularity measurements . . . . .	231



## List of Tables

1	Price of computer storage . . . . .	37
2	Historical collection University Library Basel . . . . .	52
3	Repetition code error probability . . . . .	79
4	Peviar technical specification . . . . .	87
5	Social Data from SNAP, Stanford . . . . .	114
6	Enron corpus versions . . . . .	120
7	Custodians and Non-Custodians in Enron corpus . . . . .	122
8	Unique Email Addresses of External Domains . . . . .	124
9	Multiple Email Accounts for Custodian . . . . .	140
10	List of simulation parameters . . . . .	146
11	Taxonomy of AR technology components . . . . .	170
12	Image square to aperture correspondence . . . . .	230



## A Appendix

### A.1 SFR Measurement

The SFR measurement was conducted according to the ISO standard (see [76]). As has been stated in section 2.4.2, several series of SFR measurements had to be conducted to evaluate the individual components, which by default are measured in conjunction. This part of the appendix describes only the procedure for an individual SFR measurement. It consists of three parts: image capture, image evaluation, and measurement visualization. The procedure is described in an instructional style.

#### (1) Image capture

1. Prepare the film target for capture by either enclosing it between two glass plates and sealing it with epoxy, or by putting it on a glass plate and leveling it with weights. The glass plate must have appropriate dimensions for holding by the microscope specimen holder (45-120mm by 60-90mm).
2. Prepare the microscope for Köhler illumination and place the target under the microscope and in the focal plane
3. Start the camera. The Zeiss AxioCam is set to the 1388 by 1040 scanned color mode, i.e. we have the full color resolution (no interpolation). Now, determine the measurement parameters relevant to capture, i.e. the power of illumination through the microscope lamp (no over- or underexposure), white balance
4. The parameters determined in the previous step are entered into a measurement protocol, in addition, also the time, the temperature, the relative humidity are noted
5. A measurement series is performed. A vertical or horizontal edge (depending whether one wants the horizontal or vertical SFR, which are not necessarily identical) is taken into focus, first with a black to white transition. It is captured four times. Then, an edge with a white to black transition is taken into focus, and is also captured four times. We now have eight images, which are named `sw1.tif`, `sw2.tif`, `sw3.tif`, `sw4.tif`, `ws1.tif`, `ws2.tif`, `ws3.tif`, `ws4.tif`. We call it a measurement set
6. A script is run in order to determine whether the edges are in fact rotated around  $5^\circ$ , as suggested by the ISO standard. An error of up to 1% was considered tolerable, although the standard has not given a recommendation in that regard. The following script was used to check the measurement set (this is only the code for horizontal edges):

```

% read image, convert to black and white,
% cut off top/bottom borders
img = imread(char(files(i)));
ig = rgb2gray(img);
t = graythresh(ig);
bw = im2bw(ig, t);
bw = bw(200:800,:);

% determine whether it is a black to white
% or white to black transition
compare = 0;
if(mean(bw(1,:)) < 0.5)
    compare = 1;
end

% find transition points on left and on right
% calculate angle in triangle
h1 = find(bw==compare, 1);
bw2 = fliplr(bw);
h2 = find(bw2==compare,1);
b = size(bw,2); % width of image
a = abs(h1-h2);
c = sqrt(a^2 + b^2); % hypotenuse
sinAngle = a / c;
angle = asin(sinAngle) * 180/pi;

% display message
msg = '-FAIL';
if(abs(5-angle)<1)
    msg = ' ok ';
end

```

**(2) Evaluation** Once all edges of the measurement set are in order, the set is evaluated using the `sformat2` tool. `Sformat2` is a Matlab program provided by Peter Burns, co-developer of the ISO 12233 standard. The algorithm used to determine the SFR follows the ISO standard, in which an informative C algorithm implementation is provided (see [76], p. 17). It should be noted that the `sformat2` program is not a part of the ISO standard, and is not referenced therein. For each of the measurement of the set, `sformat2` is executed, which produces the actual SFR measurement. The program has the following steps:

1. The target image is selected
2. The spatial relation between the pixels of the image and the physical size can be specified, as well as the luminance weights for the three color channels. This step was not performed in our measurements (spatial relation was applied later)
3. The region of interest (ROI) is selected, i.e. the user is prompted to mark a rectangular area containing the 5 edge. This is done manually

4. A file describing the opto-electronic conversion function (OECF) of the camera used for capture in the form of a look-up table can be selected. The OECF of our camera was evaluated and provided as a file (the camera is nearly linear)
5. The SFR calculation is performed according to the algorithm suggested by ISO. The result can be saved as a file, which is named as the input image file, although with the suffix `.fig`

Once the whole set has been evaluated, the resulting measurements are analyzed in terms of outliers. It is possible that some region of interest selection result in an abnormal behavior of the algorithm, which is easily detectable by comparing the individual measurements and noting abnormally diverging results. If the analysis yields no anomalies, the measurements can be processed and visualized.

**(3) Processing and Visualization** The resulting measurements provide the spatial frequency response in cycles per pixel. Since we want to have the measure for cycles per millimeter, the original measurement must be multiplied by the number of pixels per centimeter (which was 2.410 for our camera and microscope setup). Also, since the spatial frequencies for which the response is calculated are not identical in every measurement, the measurements must be normalized to the frequency axis of one single measurement (we have chosen the first one).

```
% the two factors are for two different lenses, we have used the
% one with factor 25
factor25 = 2410;
factor40 = 3865;
factor = factor25;

[Selection, ok] = listdlg('PromptString', 'Select Magnification factor',
                        'ListString', ['25x'; '40x'], 'ListSize', [120 120]);
if(ok>0)
    if(Selection==2)
        factor = factor40;
    end
elseif(ok==0)
    display('User canceled factorSelection, exiting.');
```

```
return;
end

% higher frequencies than the one specified here are not included
% in the report
borderFrequency = str2num(char(inputdlg(
    'Maximum frequency for local plots', 'Max.Frequency', 1, {'250'})));
if(numel(borderFrequency)==0)
    display('User canceled borderFrequencySelection, exiting.');
```

```
return;
```

```

end

%% PROCESS BLOCK - iterate all directories (multiple measurements can
%% be processed with this script)
for outerI=1:numel(dirlist)

    curDir = char(dirlist(outerI));
    curDirLabel = curDir(1:numel(curDir)-1);
    curDirText = strrep(curDirLabel, '-', '-');
    display(['processing directory ' curDirLabel]);
    display('      interpolating data...');

    % load sfr data from data files
    sw1 = load([curDir 'sw1.fig']);
    sw2 = load([curDir 'sw2.fig']);
    sw3 = load([curDir 'sw3.fig']);
    sw4 = load([curDir 'sw4.fig']);
    ws1 = load([curDir 'ws1.fig']);
    ws2 = load([curDir 'ws2.fig']);
    ws3 = load([curDir 'ws3.fig']);
    ws4 = load([curDir 'ws4.fig']);

    % create matrices that will hold the interpolated data
    % they all must have the same dimension as the matrix that
    % specifies the X axis for the interpolation, i.e. sw1
    % 'i' stands for interpolated, which will happen later
    isw1 = sw1;
    isw2 = sw1;
    isw3 = sw1;
    isw4 = sw1;
    iws1 = sw1;
    iws2 = sw1;
    iws3 = sw1;
    iws4 = sw1;

    xi = sw1(:,1); % target X Axis

    % interpolate data for all measurements so that they
    % are normalized along the frequency axis of the first
    % measurement
    isw1(:,2:5) = interp1(sw1(:,1),sw1(:,2:5), xi);
    isw2(:,2:5) = interp1(sw2(:,1),sw2(:,2:5), xi);
    isw3(:,2:5) = interp1(sw3(:,1),sw3(:,2:5), xi);
    isw4(:,2:5) = interp1(sw4(:,1),sw4(:,2:5), xi);
    iws1(:,2:5) = interp1(ws1(:,1),ws1(:,2:5), xi);
    iws2(:,2:5) = interp1(ws2(:,1),ws2(:,2:5), xi);
    iws3(:,2:5) = interp1(ws3(:,1),ws3(:,2:5), xi);
    iws4(:,2:5) = interp1(ws4(:,1),ws4(:,2:5), xi);

    % build mean
    averaged = isw1 + isw2 + isw3 + isw4 + iws1 + iws2 + iws3 + iws4;
    averaged = averaged / 8;

```

```
% calculate absolute (cycles/mm on target) frequency axis values
% note: factor depends on objective magnification, 2410 = objective 25x
averaged2 = averaged;

for i=1:size(averaged,1)
    averaged(i,1) = averaged(i,1) * factor;
end

% the average of the measurements, 'averaged', can
% now be plotted.

% ends the iteration of all directories
end
```

## A.2 RMS Measurement

The target used to determine the granularity of the microfilm consists of 52 squares of various optical densities<sup>59</sup> (shades of gray) Each square is a uniformly exposed area. Each square was captured twice, and the average image was calculated. This averaged image was the basis for the RMS granularity calculation.

The calculation happens in two steps. First, of every image, squares are cut out in horizontal direction. When cutting out squares with a side length of 20 pixels, a total of 50 squares can be cut out of the images, which have a horizontal resolution of 1000 pixels. For each of these squares, the luminance is averaged arithmetically. Of all these averaged luminance values, the arithmetic average is taken again. The margin between the average luminance of each square in relation to the average luminance of all squares is used to build the root mean square (RMS). This quadratic mean has been calculated for all optical densities.

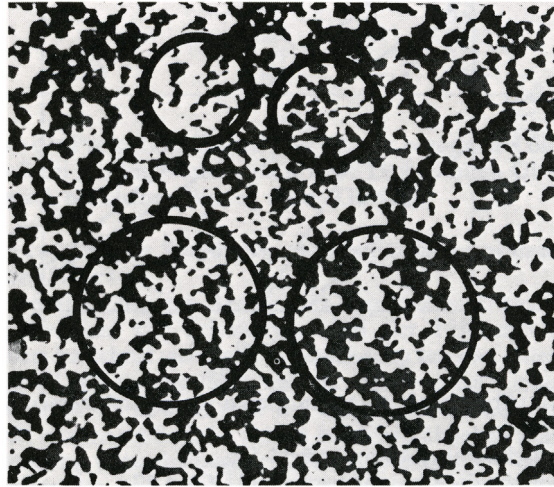
The side length of the squares that are cut out of the image are comparable to the size of the aperture in the standard Kodak Drill measurement. The side length of one pixel in the image is approximately  $0.41\mu m$ . The RMS granularity was calculated using side lengths of 1, 6, 10, 12, 16, 18, 20, 40 and 100 pixels. Table 12 compares the side lengths to their respective aperture.

Square side length (px)	Corresponding aperture ( $\mu m$ )
1	0.41
6	2.48
10	4.14
12	4.97
16	6.63
18	7.45
20	8.28
40	16.6
100	41.4
Standard Kodak Aperture	48.0

Table 12: Image square pixel count with corresponding aperture values in  $\mu m$  (Table: F. Müller)

<sup>59</sup>Optical densities, determined densitometrically, in ascending order: 0.13, 0.14, 0.15, 0.16, 0.19, 0.22, 0.23, 0.25, 0.26, 0.28, 0.32, 0.34, 0.36, 0.38, 0.4, 0.42, 0.46, 0.5, 0.52, 0.55, 0.58, 0.59, 0.64, 0.68, 0.71, 0.74, 0.77, 0.8, 0.86, 0.91, 0.94, 0.98, 1.03, 1.08, 1.15, 1.22, 1.27, 1.32, 1.39, 1.47, 1.57, 1.66, 1.74, 1.84, 1.98, 2.1, 2.11, 2.11, 2.11, 2.29, 2.28, 2.29





*Figure 85:* The effect of the aperture on granularity measurements. Very small areas can result in very high granularity (Image: H. Frieser, [193], p. 288)

Figure 85 illustrates the effect of different apertures. The Matlab script used to calculate the RMS is listed on the next page.

```

% used to locate sampling areas
startX = str2num(char(answer(1,:)));
stopX = str2num(char(answer(2,:)));
offsetY = str2num(char(answer(3,:)));
blocksize = str2num(char(answer(4,:)));

% is the image RGB (3) or grayscale (1)?
channels = str2num(char(answer(5,:)));
outfile = 'rmsmeasure.fig';
display(['Number of Patches: ' num2str(numel(files)/2)]);
display(['Blocks per patch: ' num2str((stopX-startX)/blocksize)]);
display(['Output file name: ' outfile]);

% delete any previous measurements
clear imgA imgB avg diff quad rms RMSLIST
samplecount = 0;

% for all files in the set, i.e. for all patches
for i=1:2:numel(files)
    samplecount = samplecount + 1;
    imgA = imread(char(files(i)));
    imgB = imread(char(files(i+1)));
    IMG = (imgA+imgB) / 2;

    blockcount = 0;
    for j=startX:blocksize:stopX
        try
            myMean = mean(mean
                (mean(IMG(offsetY:offsetY+blocksize,
                    j:j+blocksize,channels)))));
            blockcount = blockcount + 1;
            blockmeans(blockcount) = myMean;
        catch
            display(['Failed to create mean of block']);
            display(['BLOCK: ' num2str(offsetY) ':'
                num2str(offsetY+blocksize) ',' num2str(j) ':'
                num2str(j+blocksize)]);
            return
        end
    end
    avg = mean(blockmeans);
    diff = blockmeans - avg;
    quad = sum(diff.^2);
    rms = sqrt(quad/blockcount);
    display(['RMS for group i-' num2str(i) ': '
        num2str(rms) ' (avg:' num2str(avg) ')']);
    RMSLIST(samplecount) = rms;
end

save(outfile, 'RMSLIST', '-ascii');
plot(1:numel(RMSLIST),RMSLIST,'-ks');
title(outfile);

```

### A.3 Enron Employee List

eid	firstName	lastName	Position	Department	Location
7	Michelle	Lokay	Account Directory TW	Enron Transportation Services	
15	Monika	Causholli	Analyst	Enron North America	ENA Real Time
145	Robin	Rodrigue	Analyst	Enron Capital and Trade	Houston
12	Susan	Scott	Analyst	Enron Transportation Services	
85	Geir	Solberg	Analyst	Enron North America	ENA Real Time
135	Bill	Williams	Analyst	Enron North America	ENA Real Time
150	Liz	Taylor	Assistant to President	Enron Capital and Trade	
90	Cooper	Richey	Associate	Enron North America	ENA Fundamental Analysis
46	Stacy	Dickson	Attorney	Enron North America	
78	Gerald	Nemec	Attorney	Enron Capital and Trade	Houston
43	Bill	Rapp	Attorney	Enron Energy Services	
63	John	Lavorato	CEO	Enron North America	
28	Stanley	Horton	Chairman & CEO	Enron Transportation Services	
134	Kenneth	Lay	Chairman & CEO	Corporate	
61	Steven	Kean	Chief of Staff	Enron Energy Services	Houston
126	Rick	Buy	Chief Risk Officer	Enron Capital and Trade	Houston
125	Sally	Beck	COO	Enron Capital and Trade	Houston
124	Robert	Benson	Director	Corporate	
18	Lynn	Blair	Director	Customer Services	
123	Sandra	Brawner	Director	Enron Capital and Trade	Houston
42	Frank	Ermis	Director	Enron Capital and Trade	
39	Keith	Holst	Director	Enron Capital and Trade	Houston
107	Mike	Maggi	Director	Corporate	
105	Larry	May	Director	Corporate	
59	Brad	Mckay	Director	Enron Capital and Trade	
104	Jonathan	Mckay	Director	Enron Capital and Trade	California
91	Geoff	Storey	Director	Enron Capital and Trade	Houston
6	Kevin	Hyatt	Director Asset Dev.	Enron Transportation Services	
130	Sean	Crandall	Director Cash	Enron North America	ENA Northwest
25	Jeff	Dasovich	Director Gov. Aff.	Enron Energy Services (NA)	San Francisco
110	Andrew	Lewis	Director Trading	Enron Capital and Trade	Houston
122	Mike	Carson	Employee	Corporate	
38	Lindy	Donoho	Employee	Enron Transportation Services	
72	Tom	Donohoe	Employee	Enron Capital and Trade	Houston
121	Chris	Dorland	Employee	Enron Capital and Trade	Houston
151	Rosalee	Fleming	Employee	Corporate	
142	Dan	Hyvl	Employee	Enron Capital and Trade	Houston
70	Peter	Keavey	Employee	Enron Capital and Trade	Houston
137	Kam	Keiser	Employee	Enron Capital and Trade	Houston
32	Paul	Lucci	Employee	Enron North America	
62	Kay	Mann	Employee	Corporate	
103	Errol	McLaughlin	Employee	Corporate	
101	Patrice	Mims	Employee	Enron Capital and Trade	Houston
36	Richard	Ring	Employee	Enron Energy Services	
35	Theresa	Staab	Employee	Corporate	
44	Benjamin	Rogers	Employee Associate	Enron Capital and Trade	
20	David	Delainey	Head	Enron North America	
14	Michelle	Cash	Labor Attorney	Enron North America	
77	James	Derrick	Lawyer	Corporate	
141	Mary	Hain	Lega Specialist	Enron Capital and Trade	Houston
147	Carol	Clair	Legal Specialist	Enron Capital and Trade	Houston
79	Debra	Perlingiere	Legal Specialist	Enron Capital and Trade	Houston
29	Elizabeth	Sager	Legal Specialist	Enron Capital and Trade	Houston
120	Daren	Farmer	Logistics Manager	Enron Capital and Trade	Houston
55	Martin	Cuilla	Manager	Enron Capital and Trade	
24	John	Forney	Manager	Enron North America	ENA Real Time
53	Randall	Gay	Manager	Enron Capital and Trade	
117	Doug	Gilbert-smith	Manager	Corporate	
112	Jeff	King	Manager	Corporate	
33	Kim	Ward	Manager	Enron North America	ENA Middle Market
119	Mark	Fischer	Manager Cash	Enron Capital and Trade	Portland
133	Diana	Scholtes	Manager Cash	Enron North America	ENA Northwest
1	Robert	Badeer	Manager Team	Enron North America	ENA California
97	Matt	Motley	Manager Team	Enron North America	ENA Southwest
76	Mark	McConnell	Manager TW	Enron Capital and Trade	
82	Stacey	White	Manager, Trader	Enron Capital and Trade	Houston
34	Phillip	Allen	Managing Director	Enron Capital and Trade	Houston
113	John	Hodge	Managing Director	Corporate	
73	Mark	Haedicke	Managing Director Legal	Enron Capital and Trade	Houston

*continued on next page*

eid	firstName	lastName	Position	Department	Location
114	John	Griffith	Managing Director UK	Corporate	
66	Susan	Bailey	Paralegal	Enron North America	
3	Stephanie	Panus	Paralegal	Enron Wholesale Services	
111	Louise	Kitchen	President	Enron Online	
48	Jeffrey	Shankman	President	Enron Global Markets	
139	Jeffrey	Skilling	President & COO	Corporate	
21	Greg	Whalley	President & COO	Enron Wholesale Services	
37	Teb	Lokey	Regulatory Affairs Manager	Florida Gas Transmission	
30	Vince	Kaminski	Risk Analytics and Control	Enron Capital and Trade	Houston
26	Lysa	Akin	Senior Adm.Ass. Gov.Aff.	Enron Capital and Trade	Portland
11	Marie	Heard	Senior Legal Specialist	Enron North America	
60	Tana	Jones	Senior Legal Specialist	Enron Capital and Trade	
64	Paul	Barbo	Senior Manager	Enron North America	
144	Steven	Harris	Senior Manager	Enron Transportation Services	
108	Phillip	Love	Senior Manager Trading	Enron Capital and Trade	Houston
56	Larry	Campbell	Senior Specialist	Enron Transportation Services	
148	Chris	Stokley	Senior Specialist	Enron North America	ENA Volume Management
115	Lisa	Gang	Senior Specialist Cash	Enron North America	ENA Southwest
93	Phillip	Platter	Senior Specialist Cash	Enron North America	ENA California
45	Juan	Hernandez	Senior Specialist Logistics	Corporate	
10	Rod	Hayslett	Senior VP, CFO	Enron Global Services	
87	Holden	Salisbury	Specialist	Enron North America	ENA Real Time
86	Cara	Semperger	Specialist	Enron North America	ENA Northwest
131	Ryan	Slinger	Specialist	Enron North America	ENA Real Time
138	Mark	Guzman	Specialist Term	Enron North America	ENA Northwest
128	Eric	Bass	Trader	Enron Capital and Trade	Houston
127	Don	Baughman	Trader	Enron Capital and Trade	Houston
140	Craig	Dean	Trader	Enron Capital and Trade	Portland
118	Chris	Germany	Trader	Enron Capital and Trade	Houston
116	Darron	Giron	Trader	Enron Capital and Trade	Houston
71	Scott	Hendrickson	Trader	Enron Capital and Trade	Houston
2	Williams	Jason	Trader	Enron Capital and Trade	Houston
52	Tori	Kuykendall	Trader	Enron Capital and Trade	
69	Matthew	Lenhart	Trader	Enron Capital and Trade	Houston
143	Eric	Linder	Trader	Enron Capital and Trade	Portland
75	Smith	Matt	Trader	Enron North America	
102	Albert	Meyers	Trader	Enron Capital and Trade	Portland
100	Scott	Neal	Trader	Enron Capital and Trade	Houston
99	Joe	Parks	Trader	Corporate	
98	Vladi	Pimenov	Trader	Enron North America	
40	Joe	Quenet	Trader	Enron North America	
95	Dutch	Quigley	Trader	Enron Capital and Trade	Houston
41	Jay	Reitmeyer	Trader	Enron Capital and Trade	
94	Andrea	Ring	Trader	Enron Capital and Trade	Houston
68	Kevin	Ruscitti	Trader	Enron Capital and Trade	Houston
92	Eric	Saibi	Trader	Corporate	
67	Monique	Sanchez	Trader	Enron Capital and Trade	Houston
65	Jim	Schwieger	Trader	Enron Capital and Trade	Houston
129	Hunter	Shively	Trader	Enron Capital and Trade	Houston
146	Steven	South	Trader	Enron Capital and Trade	Houston
149	Kate	Symes	Trader	Enron Capital and Trade	Portland
88	Paul	Thomas	Trader	Enron North America	
84	Judy	Townsend	Trader	Enron Capital and Trade	Houston
83	Charles	Weldon	Trader	Enron Capital and Trade	Houston
136	Susan	Pereira	Trader (Gas)	Enron North America	
9	Jason	Wolfe	Trading Analyst	Enron North America	
19	Mark	Taylor	VP,General Counsel	Enron Wholesale Services	
132	Mike	Swerzbin	Vice President Term	Enron North America	ENA Northwest
58	Harry	Arora	VP	Enron Capital and Trade	
31	Mike	Grigsby	VP	Enron Capital and Trade	Houston
57	Arnold	John	VP	Enron Capital and Trade	
106	Thomas	Martin	VP	Enron Capital and Trade	Houston
96	Kevin	Presto	VP	Enron Capital and Trade	
74	Sara	Shackleton	VP	Enron Capital and Trade	Houston
22	James	Steffes	VP	Enron Energy Services (NA)	Houston
47	Joe	Stepenovitch	VP	Energy Marketing and Trading	
89	Fletcher	Sturm	VP	Enron Capital and Trade	Houston
50	Barry	Tycholiz	VP	Enron Capital and Trade	California
81	Andy	Zipper	VP	Enron Online	
80	John	Zufferli	VP	Enron Capital and Trade	California
17	Shelley	Corman	VP Gas Logistics	Enron Transportation Services	
51	Dana	Davis	VP Trading	Enron Capital and Trade	

continued on next page

eid	firstName	lastName	Position	Department	Location
49	Jane	Tholt	VP, Natural gas Trader	Enron Capital and Trade	
5	Drew	Fossum		Enron Transportation Services	
27	Tracy	Geaccone		Enron Transportation Services	
4	Steven	Harris		Enron Transportation Services	
109	Danny	McCarty		Enron Transportation Services	
13	Richard	Sanders		Enron North America	
23	Darrell	Schoolcraft		Enron Transportation Services	
16	Richard	Shapiro		Enron Energy Services	
54	Kimberly	Watson		Enron Transportation Services	
8	Mark	Whitt		Enron North America	

Sources for this table: [143], [144], [135].