Information Services & Use 36 (2016) 203–210
DOI 10.3233/ISU-160818
IOS Press

Why and how to avoid complex non-free software in Digital Humanities projects

Bernd Kulawik
Stiftung Bibliothek Werner Oechslin, Luegeten 11, CH-8840 Einsiedeln, Switzerland
E-mail: bernd.kulawik@bibliothek-oechslin.ch

Abstract. The 'fathers of the Internet' and developers of the TCP/IP, Vint Cerf and Robert Kahn, recently warned computer users to take care of their digital documents. Because there is no solution yet for their long-time preservation, Cerf formulated his warning as: "If there are photos you really care about, print them out." As Google's Vice President he knows what he is talking about and is working on a solution called "digital vellum": It shall be able to emulate document formats, software, operating systems and computer hardware. This paper discusses some problems with this suggestion in the light of the technical development especially from the point of view of the Digital Humanities: Obviously and – thanks to the work of Cerf and Kahn – the typical document (and the system of paradigms it is based on) is fading away, as well as dedicated software or clearly distinguishable computers used to work with it. For a short-term solution, the consequence can only be to avoid complicated and non-free software solutions, even though they may have beautiful and easily usable advantages. For a long-term solution, a basic, institutional shift in the fundamental paradigms of scientific research and its documentation is needed.

Keywords: Long-term preservation, data protection, digital dark age, information black hole

1. Introduction

> "If you have photos you really care about, print them out." [1]

With these words Vint Cerf, one of the "fathers of the Internet" and, more specifically, inventors of the Internet's most important transmission protocol, the TCP/IP, and today Vice President and "Chief Internet Evangelist" at Google, warned his audience – and every computer user – at the Annual Meeting of the American Association for the Advancement of Science at San Jose in February 2015. Cerf specifically warned of the "bit rot" when he spoke about the obvious problems of digitisation and of a sustainable and future-proof storage for digital documents in the broadest sense – be they text documents, e-mails, images, movies or anything alike:

> "We are nonchalantly throwing all of our data into what could become an information black hole without realising it. We digitise things because we think we will preserve them, but what we don't understand is that unless we take other steps, those digital versions may not be any better, and may even be worse, than the artefacts that we digitised." [1]

Cerf recently was joined by his "co-father" of the Internet, Robert Kahn, in a speech and interview Kahn gave at the IPRES conference on long-time digital preservation in Berne, Switzerland [2].

With the so-called "digital vellum" Cerf proposes a system that shall provide a solution to the problem of long-term preservation of digital data. It will consist of some sort of digital emulation of the entire computational infrastructure we use to create and work with our documents: the file formats and their descriptions and definitions, the dedicated software programs like office or e-mail programs used to create and edit these files, the operating systems on which they run in that special version used to create the documents and even the hardware on which the operating system runs.

Even if we leave aside the problems connected with the legal aspects of licensing all of these parts "for eternity" and the copy- and other rights to use them in such an emulation – and even if this would be possible at all while today some software already protects itself from being run in virtual machines: There still seem to remain some important problems – e.g., such an emulation may only lift the basic problems we are facing with the preservation of digital documents already and in the near future to a higher level. And there, on this "meta level" we – inevitably – may encounter the same problems again. Only the "documents" we will have to store and reconstruct will be larger, i.e.: entire computational systems *and* their specific states at any given point in time separated from the next by an arbitrary time span.

We do not even have to think about computer technologies of the near future like quantum computers that will *not* be digital in a narrow sense, e.g. will not calculate with ones and zeros anymore. Surely, such super-computers should be able to simulate a virtual environment for the machines and software from our digital 'stone age' without any problems. We may think even more ahead into the future and of systems that would not 'calculate' with programs and some sort of bits and bytes but, for instance, rather be comparable to biological systems like our brain. Even inside of those the simulation of a specific hardware, software (with its licenses) and documents with their admission rights etc. should be possible – except in those very common cases, where the software needs to "phone home" to get permission from a company's system to be installed and executed. Could we trust that our then historic office documents, image editing programs or database software will reach its "mothership" some 50, 100 or 200 years from now? – But that is only one, and maybe only the smaller side of the problem.

2. The disappearance of today's IT paradigms

2.1. The disappearance of the document paradigm

The problem exists already today: While Cerf like many archivists is afraid of the digital future of our documents, i.e. digital files (and collections thereof) representing them and created with office programs, saved as PDFs, digital images etc., that is: documents, for which we do not have any solution that could guarantee their availability for the next 50 or even only 25 years . . . While we are still looking for a solution to *this* problem (how to save our digital documents) – this very document paradigm for digital data is already fading. Of course, also in the foreseeable future we will want to keep pictures (or any other form of digital representation) of our beloved ones or important events, we will want to keep movies or films of professional interest like documentations or showing personal moments with our families and we will surely also want to listen to great performances of music from any time and place . . . and especially: Archives and museums will want and need to keep digital versions of historically important documents of any kind for the longest future possible. We may even suppose that any databases could be seen as a large "documents" (think of CERN's terabytes of experimental data) that could be saved in the "digital vellum", and that we will be able to preserve the (constantly changing!) states of these database file(s). Maybe there will even be a possibility not only to read but to work with these databases because the software used to manage the database and to interact with the data would also be part of the "digital vellum".

But: The digital "document" itself could, can and will be – and even is already – substituted by other forms of communication, i.e. transmission of information, that do *not* follow the same paradigm – which, in itself, is only a metaphor. When computers were very large machines with small amounts of memory space and usually handled and administered with cryptic commands, the metaphorical paradigm of the desktop with folders and files containing (text) documents, pictures etc. was invented together with the graphical user interface to make them more user friendly and to help people to orient themselves through the digital 'jungle' of the file systems, software and different "drives" of these machines. Before that and for some time even into the 1990s, the common way to interact with computers was any form of command line interface. Do we have any reason to believe that the paradigm how we work with computers will *not* change as fundamentally during the next 25 years as it did during the last 25?

For instance, "tweets" or "timeline" entries – today regularly used to report e.g. live from conferences or for discussions – and their software environments may serve as examples of what I regard as the coming mainstream of our interaction with computers, leading to the displacement, if not even the replacement of what could be called the "digital document paradigm". What should the digital archivist keep in the "digital vellum" in these cases? Which kind of "document" that could be saved resembles a *tweet* and the following comments, *re-tweets* etc.? Where do we set the limit for these kinds of responses and conversations? After one year or two? After 10 answers or 10,000? Or should we trust that Twitter, the company, will keep them for eternity? In the same way as the Usenet was "saved" or rather: not saved? The question is: What kind of different forms of communication will join or even replace the traditional document and how could these forms be saved? And who should be responsible for archiving such ever-changing "streams" of information?

In some cases, the results of such "streams" e.g. in online discussions of research topics, may finally end up as documents like it has been done for centuries. But there would still be the problem with "data streams" whose bits and parts are aggregated ad-hoc from very different sources, saved in very different file systems and computers or severs usually already distributed all over the world, literally. Where is *the* "digital document" – if there is one at all – that should be preserved for the future in these cases? While many of us today grew up with the "document paradigm", the younger generation did not. And, I guess, there are good reasons to expect that their way of using handheld computers to communicate with each other and send out or receive information in forms that do not fit into this document paradigm will prevail. They will expect data to be presented in forms compatible with their customs – and will avoid others. And museums, universities, maybe even archives are already joining this movement: In the same way an institution was not recognised as "existing" some years ago if it did not have a home page on the world wide web and could be found via search engines, today it seems to be a requirement to have a very active Facebook page with constantly changing news and discussions. My guess is that this kind of representation of content will grow and extend its reach rapidly: Who will be responsible to preserve the data (and the work and the money invested) in and for the future? And how should this be done? As far as I understand the "digital vellum", it seems not be able to provide a solution to this problem.

Of course, one may say, these forms of a somewhat "fluid" communication are comparable to talks and ephemeral contacts in the past and that we should archive only the results (or what we may regard as stopping points) in some kind of a "cutout" from these streams of information and data like the "document" that was preserved in archives or museums over millenia.

I expect the "digital document paradigm" as part of the "desktop–folder–document paradigm" to lose ground to the "stream-like" forms of communication – and it is already doing so rapidly. But still we do not even have a solution to save digital documents other than TXT and, maybe, PDF for more than 20 years. And we do not have a solution to save the digital infrastructure we use to access content in databases, though we may be able to save the database files as some kind of “documents”. But they are rather useless without the software structure needed to read them and retrieve the information not hidden in these files themselves but in the relations among the data. Therefore, we should avoid these kind of complex data structures and software and, at least, save any important information like digitised sources and documents containing the result of our research as documents in simple and free file formats.

2.2. The disappearance of the program or “app” paradigm

Another problem arises from the disappearance of dedicated software programs. Of course, we still work with different programs to create document files with our office software, save and edit images or movies, write and send emails, “tweet” short messages or update timelines. But there seems to be a development going away from the dedicated software and, therefore, away from that paradigm “digital files are created with dedicated software” which could be called the “app” or “program paradigm”. In fact, some of the programs running on contemporary computers – and surely many of the apps on handhelds – are hardly software programs in themselves. If they are, this is rather due to the commercial aspects of the software industry than to technical limitations: If we look back to the 1990s there were already office programs like Star Office that suggested to “do everything in one place” containing text, spreadsheet and presentation programs together with e-mail clients and web browser under one uniform user interface. Today, these partial programs of a larger software suite may be distributed over different servers or simply work “in the cloud”. Which, then, would be *the* “program” or software to be saved with the help of the “digital vellum”?

But we could go back even further: When Douglas Engelbart gave his famous “mother of all presentations” in 1968, he did not “open a program” to create and save a file. In fact, he saved text he had directly written *on the screen* “*to* a file” (as he said a few times), but he could jump to any text-like element (word, link, line of code etc.) “inside” these “files” from any other file he was just working on, without opening any dedicated program to work with one of these files. – Today, if we look at the “apps” on handhelds we may guess that their development moves (back) into a similar direction: Of course, and only due to the limitations of the software business and its system of licences and fees, we still work with dedicated programs, but in the background these are rather conglomerates uniting different functions like the graphical user interface and its components shared with other programs or functions providing online connection or the ability to type and display text, to record voices or images and movies etc. My guess is that in the near future, users will not be willing anymore to accept a sequence of procedures like this one:

1. Start a special program (word processor, sound recorder, camera app . . .)!
2. Create a new file!
3. Type or draw or record something. Mark a text passage and save it with copy & paste.
4. Save the file and give it a document name!
5. Start another program, open another file, copy the saved part from the first file to the new file etc.

Rather, users will (and do) expect their computers and handhelds to somehow automatically recognise the things to do and the data to save. In fact, we do not have to tell our speaking assistants anymore:

1. Open the calendar app!
2. Make an entry for an event!
3. Save the entry in the calendar app but save the information also in a general calendar file format!
4. Open the e-mail client program!

5. Create an e-mail and paste the small file with the data for this event into it!
6. Send this e-mail to my colleagues!

We simply say: “Make a calendar entry for tomorrow 10am and send it to my colleagues John and Maria.” And we expect our “intelligent” handhelds and computers to know what to do. We do not want what programs might be the right ones to do all these different steps. In this example, we may still be able to identify these programs and their steps and, therefore, we may be able to save them in a “digital vellum”. But how long will this be the case? What is in more complex scenarios? The conclusion could only be to avoid complex programs that consist of different parts distributed over different machines and that need to interact with each other constantly. And, of course, these programs should be Free Software.

2.3. The disappearance of the “computer” and/or “server” paradigm

We are already accustomed to store our data in “the cloud” when we use handhelds or laptops. But today, these devices are more powerful than most of the “super-computers” during the beginnings of the World Wide Web some two decades ago – and they are surely more powerful than those computers from the beginning of the Internet around 1970. So: Why should these handhelds e.g. communicate via special, dedicated server computers on the Internet at all? As long as I have a stable IP address or something alike, my handheld could already be constantly online as its own server delivering my documents to the rest of the world over the Internet via the built-in mobile version of Apache, for instance.

Already today, I could have all my documents, images, photographs or movies “in the cloud”, i.e. scattered over dozens of virtual servers from one or more companies. Many of these servers surely are “virtual”, i.e. distributed over several physical servers and drives, maybe not only across server racks but across entire server farms or even across several locations all over the world. (At least, this is already technically possible and could be the usual configuration soon). I still may have all these data on my mobile phone, but if it breaks: where are my data then? In my backups at home, ok. And else? I do not belong to the young generation whose life takes place in larger parts via social networks – but where are their data that should be kept for the future? Which database serving as a backend for any sort of “files” somewhere “in the cloud” should be preserved (and how often?) to document activities, opinions or even knowledge formulated in documents for the future? My guess is, the average computer user producing data worth archiving will not (want to) care about all this technical stuff. So, who will? The IT people at our universities or museums? But they themselves are often researchers who learned how to use some more complicated software like databases and do the job usually only for a few years. This is obviously not a sustainable long-term solution, especially, because this know-how is limited to a small group of persons (often only one) and because directors or managers want the most advanced, i.e. nicely looking, software to be used. And, of course, they want to use it “in the cloud”. So, the IT guys are rarely in the position to prevent the usage of non-sustainable hard- and software. And if they are, they will not stay there for a long time because they have short-term contracts only. – Therefore, the next conclusion would be: For the foreseeable future, we should only use file formats and software that can be used on a single computer without obligatory Internet connection so that the entire “ecosystem” including the hardware can be saved for the future. It should be obvious that these requirements also require to stay away from most of those new and shiny possibilities we are tempted to use to organise and present data, because many of them will not fulfill such basic requirements for long-term preservation.

3. Free software and free or open hardware

By now it also should be obvious that the aforementioned conclusions require the use of Free Software. Of course, it could be possible to re-animate Open Source Software in the future by rewriting it – but will it be allowed? Copyrights today usually are valid for 70 years after the death of the artist. What about licences and software patents? I guess there is no Open Source Software – not to speak of closed source, commercial software – that may be re-used or re-written with the permission of the original company, their programmers or their heirs in 70 years. And we should not expect that we will not have to care for these legal problems because "no-one will care anymore then". Surely, lawyers will ...

But even if we decide for and find Free Software solutions to do everything we want or need to do with our data: What about the hardware? For a few years now, hardware companies together with major software companies have been trying to "secure" our computers from malware and other attacks by preventing to run any other operating systems, e.g. free ones, on these computers than that written by a specific commercial company. If this tendency goes on and becomes the general or obligatory one (of course, only in the interest of our best and to protect us from the evil...), in the near future we may not even be able to run free operating systems and free software to access and use our data saved in free file formats. Therefore, we should vigorously support any development of free or, at least, open hardware or any hardware that will not prevent its users from using the software they want.

4. "...print them out!"

In Vint Cerf's passage cited at the beginning, he warns us, that the only way by now and for the foreseeable future (at least as long as his "digital vellum" is not finished and available) to save the digitised information that we may regard as relevant (privately or for our work) would be to print it out. But even if we follow the conclusions drawn above and accept that there is no reliable technical solution yet to save and transfer our data – be it files and documents or entire databases – and even the software and the operating systems for a period longer than 20 or even 50 years: It could be very difficult to *print* all of the documents we saved with free software in free and simple file formats – rather, it may simply be impossible, just because of the sheer amount of data. Therefore, we should find ways to "extract" the most important information from these files and print them, of course, not only once.

But how long will these printed documents and photos survive? We do not need to think of fire or other catastrophes – just think of the "life expectancy" of the ink and the paper... As far as I know there is no experience with modern inks regarding their stability over periods of more than 20–30 years. Of course, if we manage to maintain our systems, software and files by migrating them for 50 and more years, we may "print them out" from time to time again and again to avoid this ink problem. But that does not look like a durable and sustainable solution, too, does it?

5. Conclusion

So, what should or could we do to prevent our research data, digitised sources or documents and any other form or representation like databases containing and representing our research results from disappearing in the "information black hole"? How could we prevent our era from becoming the "digital dark age"?

At least for documentation and research my suggestion for a long-term solution would be:

1. Establish an institution – let's call it: Digital National Library – that is at least as stable as our oldest libraries or museums. This institution should operate on a state level or – like in the case of the German speaking countries – on an international level and be supported "for eternity".
2. This institution should develop or consolidate a complete environment of free software solutions for the common computer tasks in research and documentation for museums, archives, universities etc.
3. The institution will have to (be able to) guarantee the future development of this software environment and its compatibility with the current hardware and the entrusted data for *very* long periods.
4. Any institution using taxpayer's money for research and documentation should be obligated to use this software and adapt it to special needs *only* in cooperation with this institution.
5. Usually, projects in the Digital Humanities last only a few years. After that, (almost) no-one is or feels responsible to keep the data available for "longer" periods like decades – let alone centuries. In such cases, the institution should be obligated to offer the storage systems and transfer the data to every version of the entire software ecosystem to keep the data available "for eternity".
6. The software should also be available to companies and private persons who may entrust their data to this institution for a (low) fee.

I know that it sounds impossible to establish such a software ecosystem with software for office purposes, (electronic) publication, databases and anything alike. But first of all: We are using and adapting such software everywhere in the Digital Humanities – and almost every time a-new or even "from scratch". This is horribly expansive and most of all: not sustainable. And secondly: Going on with the uncontrolled growth of special solutions for every new Digital Humanities project simply will condemn its results – and therefore, the time, money and working power invested – to disappear in the "information black hole" within 20–50 years. What would we think of our ancestors if the research from former centuries, let's say: up to the 1950s would simply almost completely disappear? What if we had to "phone" the original clay producers to read 4,000 years old cuneiform tablets from Mesopotamia?

That's in fact what we are doing now. It is an even far worse solution than the one proposed.

So, for a short-term solution, we should follow Cerf's suggestion and print out our data. As mentioned above, this, again, does not really seem to be durable solution, at least for anything more complex than text documents and photographs. Therefore, I would like to mention a solution that for the near future could help to solve some of the most urgent problems: During the Google Summer of Code 2016 I had the chance to mentor a small project based on a software module named "ftw.book" developed by the Berne company "4teamwork" for its enterprise solutions for companies and organisations or administrations. This software lets users create and edit, review and comment webpages in the free Content Management System "Plone" in almost the same way as it can be done in any other CMS. "Plone" is based on the Web Application Server "ZOPE", containing its own web server and an object oriented database. "ZOPE" is written in Python (with some parts written in C), and therefore easily adoptable to special needs.

4teamwork's additional module "ftw.book" now adds a new content type "book" to the CMS that allows users to order their content in folders (= "chapters") and webpages (= "pages") with illustrations, tables, index and table of contents in a form very similar to a book, while everything still remains a hierarchical folder with pages in HTML and, therefore, searchable on the Internet. But with a connection to a LaTeX system running on the same server (or somewhere else), it allows the user to create a PDF based on the book and now also on article templates provided by LaTeX. Both different software environments (Python/ZOPE as well as LaTeX) and also the operating systems they are running on (Linux and other Unixes like Mac OS X / macOS Sierra as well as Windows) are well-established, and therefore, it should be possible to derive from this a general and sustainable solution. Though one would think publishers, at least those of scientific literature, would offer similar solutions to their authors to interact with them and exchange different versions of articles and books, I do not know of any – at least none publicly available. Instead, the publications process still mainly seems to consist in sending multiple files forth and back several times between authors, editors and publishers via e-mail, i.e. not very different from how this was done in the 18th century with horses – only slower.

The described combination of any common (preferably, of course, free) operating system and free and well-established, stable software already offers not only the possibility to write books or articles online, but also to keep them up-to-date. For instance: If you are working on a catalog of a collection in a museum or on Renaissance drawings today scattered all over the world (like I do), you may offer any version of the work long before it may be "completed" (which it usually somehow never is...) to an interested audience via the Internet – and you may "print it out" any time, e.g. as a book-on-demand with a timestamp every day, if you wish. Such a system could be included in the duties of the described institution. And Libraries could copy the recent version of the PDFs to their own servers regularly.

But – at least for the foreseeable future of the next 10–20 years, it seems to be wise *not* to use any non-free software, any "bells & whistles" that may look fantastic and offer "a new and better experience" to users, and anything else that has an approximately low life-expectany. Remember Java applications running in browsers? A final note: Such an institution should also offer training in the usage of its software – and to learn to work with it should be regarded on the same level of scientific competence than a second field of specialisation. The current system is punishing those who learn to use software to enable the projects in the Digital Humanities. Their colleagues who do not care about this "computer stuff" publish articles and books during the same time and receive the merits leading to the very few full-time positions – while the "IT guys" after several years in short-lived projects end up in the scientific "limbo" or "nirvana" because they "missed" to earn the merits needed for a secure position in their field. Both, the unconscious spending of money and working power in Digital Humanities projects producing data that *will not be available* anymore after 50 or even 20 years and disappear in the "information black hole" as well as the dissipation of personal energy and lifetime through these projects are nothing else than a *grotesque* misallocation of the always far too scarce resources of money and manpower – and we should rather sooner than later or the earlier the better stop to go on with both.

References

[1] www.theguardian.com/technology/2015/feb/13/google-boss-warns-forgotten-century-email-photos-vint-cerf.

[2] http://www.ipres2016.ch and http://www.srf.ch/news/panorama/der-vater-des-internets.