IT Research Challenges in Digital Preservation

© Andreas Rauber

Vienna University of Technology, Vienna, Austria
rauber@ifs.tuwien.ac.at

Abstract

Digital Preservation (DP) has evolved into a specialized, interdisciplinary research discipline of its own, seeing significant increases in terms of research capacity, results, but also challenges. With this specialization, however, core IT know-how that is needed to tackle the significant problems that we are facing in DP is not sufficiently present within the DP research community. This paper outlines some research challenges in DP, highlighting the need for the DP community to reach out to IT research in general to jointly develop solutions. It also shows some examples of integrating other computer science disciplines such as Information Retrieval / Machine Learning, or Software Engineering, to address DP challenges, concluding with a brief overview of activities at the Department of Software Technology and Interactive Systems at the Vienna University of Technology in this domain.

1 Introduction

Digital representation of information - once perceived as the solution to all problems of its analogue counter parts in terms of stability, replication, and thus long-term availability – has turned out to be more fragile and susceptible to total loss than expected. While digital objects can be replicated without any degradation in quality, it is the encoding and the dependency of the digital representation to be interpreted / rendered that is endangering accessibility. Any digital object, be it a document, a data file, or an application, requires some specific software application, such as an editor capable of interpreting ASCII/Unicode encodings, an office suite, or a database system etc., to be opened and rendered. These, in turn, rely on specific libraries, and a specific operating system, which, in turn, relies on a specific hardware environment to run. If any of these modules in the so-called view-path of a digital object is lost or defunct, the whole digital object is usually reduced to a meaningless bit-stream. Given the speed of evolution of file formats, versions of software applications, operating systems as well as hardware components, including the current drive for higher complexity at each of these levels via distributed objects, cloud computing and mesh-ups, digital objects are facing serious threats of becoming useless bit-streams within very short periods of time. At the same time an increasing amount of essential information is being produced and stored only in digital form, putting society at whole at risk.

To mitigate these risks, digital preservation has emerged as an active research discipline, combining expertise from a range of backgrounds including experts from cultural heritage and memory institutions, legal experts, scientists and engineers from a range of disciplines working intensively with scientific data, and – last, but not least – computer science and IT experts. In the last few years, numerous approaches have been analyzed, standards have been devised, and systems are being deployed to tackle the challenges. In this short time period, Digital Preservation (DP) has evolved into a research discipline in its own right, with experts collaborating intensively in an interdisciplinary manner, successfully driving both our understanding of the problem as well as the availability of solutions. However, due to the highly interdisciplinary nature of the challenges, as well as due to its evolution into an independent research field of its own, DP research runs the risk of excluding essential expertise and input from more traditional sub-disciplines in each of the various disciplines involved. This may be due both to the natural segregation happening with any sub-group forming, as well as due to the complexity inherent in interdisciplinary research, rendering both language as well as communication forms difficult to understand for the non-initiated.

This paper tries to shed some light on the complexities and challenges in DP research that require specific involvement from a range of core computer science disciplines. The list is by no means exhaustive, nor is its focus on computer science aspects meant to downplay the other disciplines involved. It shall merely act as a call to experts in the respective disciplines to consider devoting effort to these non-trivial challenges that are within their core expertise to help advancing the field and pushing it further in order to solve a problem that may well turn out to be the greatest disaster for an information society relying on digital information and processes if leaft unsolved.

Proceedings of the 11th All-Russian Research Conference «Digital Libraries: Advanced Methods and Technologies, Digital Collections» - RCDL'2009, Petrozavodsk, Russia, 2009.

The remainder of this paper is organized as follows: The next section provides an overview of research agendas in the field of digital preservation. It also lists some of the EU-funded research projects launched recently in this domain. Section 3 collects a number of challenges in various sub-disciplines in IT, trying to motivate the need for research and potential directions. Section 4, finally, summarizes a number of efforts currently worked upon in our group. Section 5 summarizes the paper and points to some recent initiatives in terms of focussed DP education.

2 Research Agendas in Digital Preservation

Due to the pressing importance, a number of research agendas for Digital Preservation have been compiled in the last few years, and a series of research projects have been launched in Europe and all over the world to tackle these issues.

One of the most recent research agendas is the DPE Research Roadmap (DPE Project Consortium, 2007).

It is based on the analysis of a number of earlier research roadmaps in this domain, and – while acknowledging advances in certain domains such as specifically the creation of conceptual models and a common understanding of the problem domain – emphasizes the needs for interoperability and further standardization in numerous areas. Specifically, it recommends research in the domains of restoration and conservation; risk analysis and mitigation; understanding and handling of significant properties of digital objects and their context; interoperability amongst systems and automation of workflows, up to challenges in the general management of preservation activities, systems and organizations. Storage systems still pose numerous challenges. Last, but not least, a strong emphasis is placed on the need for controlled experimentation and evaluation to obtain a scientifically valid basis for decisions. It also provides an extensive review of earlier projects and national as well as international activities at that time.

Based on the need for further research and development, a range of activities was started, specifically within the 6th and 7th Framework Programmes of the European Commission. Two large initiatives started already within the 6th Framework Programme are the integrated projects Planets and Casper. Planets[1] (Preservation and Long-term Access through Networked Services) is a four-year project. It aims at building practical services and tools to help ensure long-term access to our digital cultural and scientific assets. It consists of a range of services embedded within an interoperability framework, comprising Preservation Actions, Preservation Characterization, a solid Preservation Planning workflow (H. Kulovits and A. Rauber, 2008), as well as a Testbed for service evaluation. CASPAR[2] (Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval**)** is an Integrated Project aiming at implementing, extending, and validating the OAIS reference model (ISO14721:2003). It intends to design virtualisation services supporting long term digital resource preservation, despite changes in the underlying computing (hardware and software) and storage systems, and the designated communities, as well as to integrate digital rights management, authentication, and accreditation as standard features.

Under the 7th Framework Programme, a number of new initiatives have been launched[3], including specifically Keep[4], a project focussing on emulation as the core preservation strategy; PrestoPrime[5], focussing on the preservation of audio-visual media in the broadcasting domain; LiWA[6], tackling the challenges of long-term preservation of data in web archives, with a specific focus on the evolution of semantics across time; PROTAGE[7], which is relying on agent-based environments for the preservation of digital objects; as well as SHAMAN[8], which will implement large-scale European-wide collections with access services that support communities of practice in the creation, interpretation and use of cultural and scientific content, including multi-format and multi-source digital objects.

The progress that these projects make, as well as the increasingly challenging scenarios that they address, are good indicators of a maturing research discipline. While it is definitely a good direction for the community to establish itself, it also brings along some of the downsides of specialized sub-disciplines. Amongst them are increasingly closed circles of highly specialized experts working in the field, using specialized terminology, and building upon a large body of pre-existing knowledge that makes it hard for other disciplines to contribute in spite of the fact that specialized contributions from other sub-disciplines would be highly beneficial. One particular challenge, inherent in the interdisciplinarity of the field, stems from the fact that the most active communities for a long time came from the cultural heritage and archival sciences domains. As a consequence, we have found it harder to attract the interest of IT experts in different fields to address the challenges in digital preservation, and to understand the impact that their respective areas of expertise can have on this community.

One notable initiative to bridge this gap is the DPE Digital Preservation Challenge[9]. This competition series consists of a number of digital preservation tasks addressed specifically at computer and information science students. Examples include the identification and recovery of information from a binary data stream, such as e.g. recovering the sound played during the opening sequence of an old computer game; retrieving the address of a specific person from an obsolete database; or developing a solution for preserving early works of digital art. Faced with challenges like these, students immediately realized the difficulty and complexity of the problems – as opposed to the rather laissez-faire approach of suggesting that simple software updates and import/export functions could overcome all problems of IT obsolescence. It also made DP challenges more accessible and exciting to address, phrasing them as actual IT challenges, mitigating part of the community jargon problems.

The following section tries to highlight – without any claims of completeness – some of the challenges in DP and their relationship to core computer science disciplines, inviting further comments and contributions.

3 Digital Preservation and IT Research

Digital Preservation, by its very nature, is tightly bound to expertise in information technology and computer science. Nevertheless, formulating requirements and challenges using terminology such as a*uthenticity, archival workflows, ingest/deposit regulations, appraisal, significant properties,* and others does not make these obvious areas for computer science research. As a result, relatively few computer science experts are actively involved in research activities in this domain.

Amongst the more obvious IT topics within DP is storage. Preservation systems need to store massive amounts of data for long periods of time with high levels of data security. While these challenges are largely in-line with the requirements of traditional IT systems, DP adds some specific challenges. One of these is the need for a high stability of the storage medium in off-line conditions, i.e. without the need to be continuously operated or monitored. Also, self-checking and verification are essential. However, even more than the stability of the storage medium, the stability, simplicity and long-term availability of reading devices for the storage media is essential. With some tape storage devices, the durability of the media exceeds the availability of the technology necessary to read/write the tapes, rendering such media unreadable not due to media decay, but to HW obsolescence. While preservation actions such as data migration (conversion to new formats), which require periodic copying of data, mitigate this risk by allowing media migration at the same time, many institutions adopt a policy of keeping a copy of the original object ingested into a preservation system permanently and unchanged, calling for specific storage technology. Hybrid storage of both digital as well as analogue representations of objects, self-maintaining hierarchical storage systems, description of storage technology, large-volume transfer of several petabytes of information, error- and consistency checking across such large amounts of data, etc. all need to be addressed from the point of view of long-term preservation systems in times where replacement cycles of technology get shorter and shorter.

One of the core concepts of DP lies in the requirement of being able to provide authentic copies of digital objects in the future. Authenticity, in a nutshell, refers to the characteristic of an object to represent what it pertains to be, i.e. to show all characteristics, both technically as well as intellectually, that allow is usage. While ultimately this is all about trust and documentation of any changes that may have happened to an object, it also requires a consistent audit trail and software solutions to ensure that an object is not altered in an unauthorized and undocumented way, be it on purpose or accidentally. This poses high challenges in terms of IT security on preservation systems, requiring documentation and enforcement of access rules across long periods of time. Key management and the security of various generations of encryption technology need to be managed across several system replacement cycles. Revocation timestamps of keys need to be documented, and a complex management of users and roles needs to be in place and consistently monitored. Ensuring that an object remains encrypted for time periods of 70 years while allowing decryption – transformation – re-encryption to happen within a sealed-off black-box poses non-trivial challenges. Trustworthiness of such systems will depend on audit routines that require solid means for automatic verification of code, workflows, and designs.

This is further complicated by the fact that any preservation system built and audited against these requirements (which already constitutes a complex task) needs to undergo a series of revisions and re-implementations throughout its "life time" of 100+ years. The levels of documentation needed to understand and validate performance without compromising security are hardly analyzed in current DP research. In fact, the very need for specialized DP systems partially comes from the fact that operational IT systems lack the long-term perspective of other areas of systems engineering. Other disciplines, such as architecture, plan the whole life cycle of a system (bridge, industrial plant, etc.), including periodic maintenance routines up to its destruction and replacement already during the design phase. Similar maintenance and replacement cycles are planned as part of the design of cars, planes, infrastructure networks, and other complex systems. IT infrastructure and software systems, on the other hand, are often planned on a rather hic-et-nunc basis, with software maintainability often being reduced to a marginal conceptual requirement, and software system succession usually not considered at all. If these requirements were considered as an integral part of system modelling and software engineering processes, would then "OAIS compliance" be an integral characteristic of all IT systems? While obviously desirable, principles for such long-term levels of software life cycle management, costing, modelling and development principles still require intense research efforts.

The long-term aspect itself poses significant challenges, as the semantics of terminology used to describe system functions, actors, or requirements is likely to change across such large time horizons. Knowledge assumed to be generally available today may be obscure in the not-too-distant future, rendering descriptions of system architectures, HW components and algorithms incomprehensible. This is especially important considering the fast innovation cycles and the resulting fashion-like adoption of new, short-lived terminology.

While a lot of current efforts in DP focus on document type objects, some of the most valuable information is stored in databases of considerable complexity. While a clean separation of data and functionality has always been a core requirement in data base design, many real-life systems do not fully comply with this desired ideal. Databases not only host information, they act on it. Semantics is hidden in distributed tables; triggers and active code add further complexity. Design and documentation guidelines for databases that are preservation-aware would ease subsequent preservation of such systems and their valuable contents.

But even the rather simple concept of "digital objects" is likely to pose drastically new challenges in the near future. Most approaches for document-type objects such as images, office documents, spreadsheets, videos, etc. are based on the concept of (complex) files following specific file format definitions. However, recent development saw more and more file types evolve into generic containers that can include virtually any other file type container, hardly without limitations. This may ultimately lead to the death of the concept of a file format as a well-defined entity which restricts the technical characteristics of the bit-stream that is embedded. It poses drastic new challenges to the tools operating on such files, including a recursive decomposition of objects if they are supposed to deal with all objects of a certain type. Additionally, mash-up documents integrating distributed on-line sources, as well as increasingly active elements within files turn them into both programs as well as a network of dynamic content, rather than static objects that can be stored and handled. This will require drastically new approaches to both migration and emulation as the two dominant preservation strategies of current times.

Understanding what a digital object consists of, both technically as well as semantically, requires sophisticated concepts from data mining as well as information retrieval. While the challenges of analyzing and describing the structure of large amounts of files are only understood to a limited degree (C. Becker et al., 2008), those of semantically searching and mining the data are more obvious. As a rather obvious example we can see the challenges of providing full-text search within a Web Archive as the equivalent of operating a current state-of-the-art search engine on an index that does not only span the current Web, but also its entire history, requiring multiple server farms for storage and index handling (even though, very likely, for lower query traffic.) This will be complemented by new search technology, specifically in the multimedia domain, to enhance traditional text search. It will go beyond simple retrieval of objects, but addresses whole networks of objects and their semantics that may well be different from the semantics of the individual objects and at different periods of time.

This leads to another interesting aspect of digital preservation, where IT research closes the loop again to social sciences and ethics. While heritage institutions and research centres have always been collecting and maintaining data over long periods of time, these data usually were complicated to search through. Utilizing these assets was a complex research endeavour in its own right. With new search technology and assuming the availability of good long-term preservation solutions, masses of historic data can be analyzed efficiently, giving rise to concerns about the value of the concept of forgetting within a society. These issues are being addressed in the fields of data protection laws for operative databases. The impact of long-term availability and aggregation of data, for example within Web Archives, and its relation to advances in search technology is still far from being fully understood (A. Rauber, M. Kaiser, and B. Wachter, 2008).

The list of topics may be continued for quite some times. Further significant challenges that are essential for mastering the long-term preservation of digital objects include automatically identifying semantics from code, cross-compilation, abstraction from hardware layers, chip design and documentation, and others.
Many of these challenges are already subject of intensive research within the respective disciplines, independent of the DP research agenda. Connecting them and adopting them as standard best-practice principles may help moving computer science and information technology to the same level of maturity that other, older disciplines and specific sub-disciplines within IT may have reached already.

4 Selected DP Research Activities

This section briefly reviews some of the activities in digital preservation in our group. We will start by analyzing preservation planning as a specific type of Commercial-off-the-Shelf (COTS) component selection in Section 4.1. This is followed by a description of automating preservation services using rule-based decisions to provide preservation systems to small office/home office (SOHO) settings as part of the HOPPLA system in Section 4.2. While emulation has always been one of the dominant preservation actions, solidly evaluating and benchmarking emulators proves more challenging than expected. In Section 4.3 we will briefly review some of these challenges. In Section 4.4 we will switch to understanding the context of creation and usage of digital objects, relying on information retrieval and data warehousing principles to automatically establish context of usage. In Section 4.5 we will take a glimpse at using microfilm as a viable storage technique for binary data, before concluding in Section 4.6 by raising some ethical issues involved in archiving and analyzing data over long periods of time, specifically in the domain of Web Archiving.

4.1 Preservation Planning as COTS Selection

A range of different strategies, i.e. preservation actions, have been proposed to tackle the digital preservation

challenge. However, which strategy to choose, and subsequently which tools to select to implement it using which system configuration and which parameter settings, is a crucial decision. It must be based on a well-documented and profound analysis of the requirements and performance of the tools taken into consideration. The Planets Preservation Planning approach (Strodl et al., 2007) allows the assessment of all kinds of preservation actions against individual requirements and the selection of the most suitable solution. It enforces the explicit definition of preservation requirements and supports the appropriate documentation and evaluation by assisting in the process of running preservation experiments. It is based on work performed in the DELOS Digital Preservation cluster, first introduced in (C. Rauch, and A. Rauber, 2004). While the workflow developed for evaluating DP solutions may seem strongly DP-centric, it is, in fact, a highly repetitive scenario similar to commercial off-the-shelf component selection. Thanks to strictly formulated requirements and the highly standardized behavior of the components to be evaluated (basically, a pair of input and output objects, and system performance being evaluated against a large set of criteria, referred to as "objectives" in the preservation planning process) migration tools and emulators can be evaluated following procedures similar to those employed in COTS selection scenarios. Vice-versa, COTS selection may generalize the evaluation principles tested in DP settings to be utilized for generic COTS selection settings. (C. Becker and A. Rauber, 2009) Current work in this direction concentrates specifically on automating the evaluation of preservation actions by providing a sophisticated measurement framework to create quality-aware web services. (C. Becker et al., 2009)

4.2 HOPPLA: Automating Digital Preservation

Digital information is of crucial value to a range of institutions, from memory institutions of all sizes, via industry and SME down to private home computers containing office documents, valuable memories, and family photographs. While professional memory institutions have dedicated expertise and resources available to care for their digital assets, SMEs and private users lack both the expertise as well as the means to perform digital preservation activities to keep their assets available and usable for the future. The Hoppla (Home Office Painless Persistent Archiving) system provides digital preservation solutions specifically for small institutions and small home/office settings. (S. Strodl et al., 2008) It hides the technical complexity of digital preservation challenges by providing automated services based on established best practice examples. Appropriate preservation strategies and required tools for performing them are delivered via a web service, effectively outsourcing the required digital preservation expertise.

4.3 Evaluating Emulators

Within the preservation planning process, a range of different preservation actions, such as specific migration tools in different configurations and with specific parameter settings are evaluated, in how far they meet the requirements set out for the specific object collection at hand. While this is a difficult task for migration tools, it becomes even more complex in emulation settings. The reason is that, usually, for migration approaches "only" the final rendering of an object needs to be evaluated, with rendering including all characteristics of an object be it visible display, but also acoustic as well as structural aspects. In emulation settings, however, an important focus lies on the interactive aspects of a digital object. Behaviour of an object needs to be evaluated not only in a static sense, but it may also include a specific timeline. Furthermore, the results may depend on external factors such as random elements popular in computer games, or networked behaviour of objects. In order to understand, whether a specific behaviour of an object is correctly preserved or whether artefacts are introduced by the environment provided by the emulator is non-trivial. These aspects are further complicated by the fact that it is not obvious, at which level to compare the result of rendering a digital object in an emulation setting. Different options available include the representation on a set of output devices such as loudspeakers and the screen, the representation on internal memory such as the graphics card, or the internal process status in main memory, accounting for different artefacts introduced by components that may be beyond the control of an emulation system, such as the screen resolution and colour representation scheme available at a specific hardware platform. While ad-hoc evaluation may be sufficient in specific settings (M. Guttenbrunner et al, 2008), more systematic approaches are required, leading to specific design guidelines for emulators in digital preservation settings (M. Guttenbrunner, 2009)

4.4 Establishing Context of Digital Objects

The context of objects is essential for the interpretation of information entities, for establishing their authenticity as well as ensuring appropriate use. Thus, documenting the context of creation and use is an essential task in digital library and document management settings, for retrieval tasks as well as for digital preservation. Yet, context is notoriously difficult and labour-some to establish and document, and often missing or partially incomplete or incorrect when it has to be entered manually by the creator of the objects.

To address this challenge we are researching methods to (semi-)automatically determine the creation and usage context of digital objects (R. Mayer and A. Rauber, 2009). Various aspects of context in different dimensions are automatically detected, and different views at multiple levels of granularity allow the extraction of the most appropriate connections to other digital objects.

Context exists in several forms, ranging from a very low-level technical context in which the object was created, via its immediate context of use (people involved, the project or activity it is related to, etc.), to a wider sociological, legal or cultural context. All levels

of context are of importance for the authentic interpretation and usage of a digital object. However, we focus predominantly on the narrower focus of context that can be determined (semi-) automatically. We thus consider the detection and documentation of context of digital objects as a semi-automatic process along several partially orthogonal dimensions, each of which structures objects according to different aspects. We currently use the following dimensions in our first prototype:

- the time of object creation and modification
- the object type
- the people involved
- the content across different sub-categories, such as (a) topic (b) genre, and (c) acronyms, for example in project names.

The concept of using various dimensions as orthogonal views on the data is inspired by the concept of data warehouses and the data analysis method used therein, on-line analytical processing (OLAP) (R. Kimball and M. Ross, 2002). A central concept is the OLAP cube, which prepares the data for fast multi-dimensional queries and analysis. The analyst can pivot the data in various ways, e.g. see all the sales for a specific city for a certain product, and do this at various levels of aggregation, allowing easily to obtain a more detailed view on demand (`drill down') or a more abstract, summarised view (`roll up').
Establishing context along these and other dimensions in combination with appropriate tools for visualizing, grouping and exporting context information supports a range of different application scenarios, such as object ingest in digital repositories, disaster recovery (R. Mayer, R. Neumayer, and A. Rauber, 2009), or user support in information retrieval tasks within archival holdings.

4.5 Storage on Microfilm

Digital data is prone to decay on several levels, one being the storage of data (bit-level preservation), ensuring that the data is securely stored on data carriers that can be read with current technology. A second layer is the logical preservation: digital objects require specific software to be opened and read, which in turn require specific operating systems, device drivers, and, ultimately, hardware to run. Finally, semantic preservation is essential to facilitate correct interpretation of objects, similar to conventional analogue pieces of information.
Several solution are being implemented, usually relying on regularly migrating data both from old storage technology to current one, as well as format migration to current versions of file formats.
However, specifically the latter usually incurs changes to the objects, some of which may seem undesirable with respect to their future usage, especially since these changes accumulate over a series of migration steps.
While careful planning procedures try to limit this effect (Strodl et al., 2007), it cannot be avoided completely. Thus, most initiatives recommend to always maintain the original format version of an object to allow reconstruction and access e.g. via emulation if required. This, in turn, calls for a cost-effective strategy for bit-level preservation of digital data on a durable storage technology not requiring regular maintenance.
Additionally, most settings require a stable back-up copy to be maintained for all data (including migrated versions) in addition to on-line versions for continuous use.
Unfortunately, most digital storage techniques do not offer themselves for these purposes: hard disk RAID arrays require regular operation and need to be replaced every few years. Tape drives as long-term storage of massive amounts of data require regular re-winding of tapes to maintain them readable.
Further, with current development cycles the durability of the tapes has surpassed the support life-time for tape readers, rendering the respective tapes unreadable unless migrated to new types of tapes.

This has led to the revival of a rather unexpected storage technique for digital data, namely microfilming (C. Voges, V. Мдrgner, and T. Fingerscheidt, 2008). Microfilm, especially black/white film, has proven a very durable media, requiring no maintenance apart from appropriate storage conditions. Microfilm as storage media has a life span of more then 100+ years and has the advantage that a media migration has to be done less frequently. It is already used for long term storage of scanned images of paper documents.
Provided correct encoding schemas are used, microfilm can store both analogue representations (i.e. images) of objects as well as the digital data stream, offering the additional benefit of easy inspection and redundancy of representation forms, as it basically already includes a kind of "migrated" analogue representation in addition to the digital object. We are thus currently investigating different encoding schemas such as UUencode or XXE as well as 2-d barcodes as means to “print” digital data to microfilm, and to recover it by using scanning and OCR technology with subsequent decoding into binary data.

4.6 Ethical Issues in Web Archive Creation & Usage

A completely different type of challenge is posed by the accumulation of information across long periods of time in combination with advances in search technology, namely questions concerning ethically correct creation as well as usage and provision of archived data (A. Rauber, M. Kaiser, and B. Wachter, 2008). This is particularly prominent in the domain of Web Archiving. While Web Archiving initiatives rescue a massive amount of information on the Web from being permanently lost, the massive collection of Web data poses not only fascinating possibilities for accessing a wealth of information, as well as an invaluable resource for scientist wanting to understand the technological and sociological development of the Web and society at large. It also constitutes a new type of information on its own, posing numerous ethical challenges, specifically given the powerful techniques for analyzing and exploring the masses of accumulated information that

we will have available in the near future.
Being aware of this issue, most Web Archives currently strictly limit access to their holdings, or provide means to allow people having their content excluded from holdings to avoid the subsequent challenges, at the same time drastically limiting their value and usefulness. While the ultimate solution to the problem of what kind of access will be permissible will have to be a legal one, it is important to understand the detailed characteristics of the possibilities of "unethical" exploitation of a Web Archive's holdings, or simply types of usage that some people may feel uncomfortable with concerning the content that they have made available on the Web sometime in the past. It requires a profound understanding of the semantic and cognitive aspects and values of the information aggregated over time, as opposed to the individual pages it is based upon and that are currently searchable via conventional Web search engines. It will also require the development of technical means to counter these challenges. This requires researching the potential of new techniques of large-scale information retrieval including its semantic capabilities and, specifically, the drastically different level of semantic information that can be gleaned from the unprecedented collection of information that is available in Web Archive holdings.

4.7 Other activities

A number of other activities are currently being explored in our group, addressing issues such as archiving non-traditional environments on the Internet, with a specific focus on Virtual Worlds such as SecondLife. While approaches to completely preserve the data structures as well as rendering environment may provide the most comprehensive solution for preserving the world as such, the interaction happening in these worlds may be at least as important to capture. To this end we are investigating automated means of filming activities in certain areas of these virtual worlds while trying not to capture personal information about avatars and their users.
Another line of activity aims at recovering information from obsolete digital objects by analyzing the structure of an object's code in order to determine regular patterns that can help in its decoding.
We are also performing first experiments to model how "forgetting" may be implemented in a digital archive, specifically in settings where multiple versions of an object are archived incrementally, each of which may be available via a range of different migration paths, analyzing information growth between versions as well as the characteristics of complementary object representations in different formats.

5 Summary

Digital Preservation represents some tremendous challenges that need to be addressed within the near future if we want to ensure that the wealth of information that we are creating continuously is to remain accessible and usable in the near and far future. This affects all levels of society and all business domains, starting from cultural heritage institutions and science data centres, via industry and business, up to small and medium enterprises as well as home users.
A growing community of researchers has evolved in the last few years that are investigating the specific challenges in digital preservation from their respective backgrounds. However, in order to really solve the challenges, a wider range of experts particularly from core computer science disciplines needs to get involved and excited about the research challenges waiting in this field, if possible making DP an integral part of all IT systems, ultimately achieving sustainable computing at all levels.

The digital preservation community has become aware of this need. Specific initiatives such as the DPE Digital Preservation Challenge have been devised to attract the interest of Computer and Information Science students and demonstrate the hard tasks waiting to be solved.
Complementing this, specific curricula are being developed to train experts in the field of digital preservation. In this tradition we find the series of summer schools organized by DELOS[10] and nestor[11], as well as dedicated curricula initiatives such as DigCCurr[12], the German initiative for a training course for professionals in Digital Preservation, as well as an emerging initiative for a European Master in this domain aiming at providing a solid education for future experts to advance the field. Digital Preservation is also represented as a major field of specialization in the Digital Library curriculum developed under an NSF grant by Ed Fox and his team at Virginia Tech[13].
While the field is still young, it has matured to a level of considerable complexity and specialization. In order to solve the challenges ahead of us, however, the preservation community needs to ensure it remains open and manages to attract professionals from different backgrounds, including but definitely not limited to, computer science experts, to jointly address the challenges that our information society is facing.
On the other hand, Computer Science has to accept the need for achieving sustainable computing, considering system operation, maintenance and replacement as an integral part of the system design and development process. Once this level of IT system maturity is reached, DP will come "for free" as part of systems operation, rather than as a separate add-on. Until then, we will need to continue investing considerable efforts to mitigate the risks threatening the long-term availability of digital information.

References

[1] C. Becker, H. Kulovits, M. Kraxner, R. Gottardi, and Andreas Rauber. (2009).
An Extensible Monitoring Framework for Measuring and Evaluating Tool Performance in a Service-oriented Architecture. In: Proceedings of the 9th International Conference on Web Engineering (ICWE 2009), LNCS 5648, Springer.

[2] Christoph Becker, Andreas Rauber. (2009). Requirements modelling and evaluation for digital preservation: A COTS selection method based on controlled experimentation. In: Proceedings of the ACM Symposium on Applied Computing (SAC'09), Track 'Requirements Engineering'. Honolulu, Hawaii, USA, March 9-12, 2009.

[3] Christoph Becker, Andreas Rauber, Volker Heydegger, Jan Schnasse, and Manfred Thaller. (2008). Systematic Characterisation of Objects in Digital Preservation: The eXtensible Characterisation Languages. Journal of Universal Computer Science, 14(18):2936-2952.

[4] DPE Project Consortium. (2007). Research Roadmap. Project Deliverable DPE-D7.2. http://www.digitalpreservationeurope.eu/publications/reports/dpe_research_roadmap_D72.pdf

[5] M. Guttenbrunner, C. Becker, A. Rauber, and C. Kehrberg. (2008). Evaluating strategies for the preservation of console video games. In Proceedings of the Fifth international Conference on Preservation of Digital Objects (iPRES 2008), 115-121.

[6] M. Guttenbrunner (2009). Evaluating the effects of emulation environments on rendering digital objects, Planets Deliverable PP/5-D2

[7] ISO14721. (2003). Space data and information transfer systems – Open archival information system – Reference model (ISO 14721:2003).

[8] R. Kimball and M. Ross. (2002).The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling (Second Edition). Wiley.

[9] H. Kulovits and A. Rauber (2008). Preservation Planning with Plato. In: Proceedings of the All Russian Conference on Digital Libraries (RCDL 2008).

[10] R. Mayer and A. Rauber. (2009). Establishing Context of Digital Objects' Creation, Content and Usage. In Proceedings of the JCDL Workshop on Innovation in Digital Preservation (InDP 2009), Austin, Texas, USA. June 19 2009.

[11] R. Mayer, R. Neumayer, and A. Rauber. Data Recovery from Distributed Personal Repositories. In: Proceedings of the 13th European Conference on Digital Libraries, ECDL 2009, LNCS, Springer. Corfu, Crete, September 2009.

[12] A. Rauber, M. Kaiser, and B. Wachter. (2008). Ethical Issues in Web Archive Creation and Usage – Towards a Research Agenda. In: Proceedings of the 8th International Web Archiving Workshop, Aalborg, Дnemark.

[13] C. Rauch, and A. Rauber. (2004). Preserving digital media: Towards a preservation solution evaluation metric. In Proceedings of the 7th International Conference on Asian Digital Libraries (ICADL 2004) (p. 203-212). Berlin, Heidelberg: Springer.

[14] S. Strodl, F. Motlik, K. Stadler, and A. Rauber. (2008). Personal & SOHO Archiving, In: Proceedings of the Joint Conference on Digital Libraries (JCDL 2008), June 16-20, 2008, Pittsburgh, Pennsylvania, USA.

[15] S. Strodl, C. Becker, R. Neumayer, and A. Rauber. (2007). How to Choose a Digital Preservation Strategy: Evaluating a Preservation Planning Procedure. In: Proceedings of the ACM IEEE Joint Conference on Digital Libraries (JCDL'07), Vancouver, British Columbia, Canada, June 18-23, 2007.

[16] C. Voges, V. Мдrgner, and T. Fingscheidt. (2008). Digital Data Storage on Microfilm - Error Correction and Storage Capacity Issues. In Proceedings of IS&T Archiving Conference, Bern, Switzerland, June 2008.

[1] Planets, http://www.planets-project.eu/
[2] CASPAR, http://www.casparpreserves.eu
[3] Digicult Projects FP7, http://cordis.europa.eu/fp7/ict/telearn-digicult/digicult-projects-fp7_en.html
[4] Keep, http://www.keep-project.eu/
[5] PrestoPRIME, http://www.prestoprime.eu/
[6] LiWA, http://www.liwa-project.eu/
[7] PROTAGE, http://www.protage.eu/
[8] SHAMAN, http://www.shaman-ip.eu/
[9] DPE Challenge http://www.digitalpreservationeurope.eu/challenge
[10] DELOS Summerschools on Digital Preservation, http://www.dpc.delos.info/ss08/
[11] nestor winter/spring/summerschools on digital preservation, http://nestor.sub.uni-goettingen.de/education/index.php?lang=en
[12] DigCCurr, http://ils.unc.edu/digccurr/index.html
[13] DL Curriculum, http://curric.dlib.vt.edu/