

A Holistic Approach to Bit Preservation

by Eld Maj-Britt Olmütz Zierau

```

                                01001
                                0100  010010101110110
                                011001 100110011100
                                1011011 01001010100010
                                101110110110
                                11010010100110
                                100100010010
                                110100 1010
                                0010 100
                                010
                                011001
                                01 11001
                                010010 010001001011
                                010010 1011011
                                101100110
                                111000110101
                                01010001010111
                                101101101101001010
                                011010 010001001011
                                0 01 00 001001011111
                                01 01 1 0100010110010
                                11 101 101011001100100100101
                                1100 01000101010011001001001
                                1100100010101001100100100111
                                110010001010100110 010 00111
                                1100100 010 10100110 01001
                                11001 0 0010101 001100100100
                                11 001 110 1000100100001
                                0010001101 10100101
                                100110010001101 101001010
                                0101000101011100101111100111100111
                                101101101101001010001011100111001111
                                011010000100101000010111111100111100
                                10001011001010101100110010010011100111
                                101010001010001100100100110011111001
                                101010001010001100100100110010101000
                                101010011001001101010001010100110
                                0100110101000101010011001010101110
                                0010101001100101010001010101
                                001010100010101101010101
                                001010100010101 0101
                                0010101000 00
                                00 10101000101 1
                                001010
                                001010 1011
                                11001010
                                10100010
                                110
                                1100111010
                                0 0010011101101
                                10011001000110101
                                01010001010111001011
                                1011011011010010100010
                                01101000010010100001011111
                                100010110010101011001100100100
                                1010100010101001100100100110011
                                101010001010100110010010011001
                                10101000101010011001001001100
                                10101000101010011001001001
                                10101000101010011001001
                                10101000101010011001001
                                10101000101010011001001
                                10101000101010011001001
                                1010100010101001100
                                1010100010101001100
                                1010100010101001
                                0101000101
                                101

```

Preface

This thesis is submitted in order to obtain the degree of Doctor of Philosophy at the Department of Computer Science, Faculty of Science, University of Copenhagen (DIKU). The PhD is submitted without prior admission. Research documented in five peer-reviewed papers forms the basis of this article-based PhD. The research has been conducted as part of my position as a digital preservation specialist at the Royal Library of Denmark.

For making this thesis possible, I would first like to thank my boss, Birgit Nordmark Henriksen, for her encouragement, and for giving me the opportunities to do the research that is the basis for the papers included in this thesis. I would also like to thank my family and friends who have supported me, especially my husband and my children whose daily life has been strongly affected by this work.

For contributions to the contents of the thesis, there are many people to thank. This has included very skilled people with different perspectives on the subject who have inspired me through discussions and reviews.

I want, of course, to thank my co-authors and my colleagues in the Digital Preservation Department of the Royal Library of Denmark. Here I would like to give special thanks to my colleague Ulla Bøgvad Kejser and my former colleague Anders Sewerin Johansen for their valuable cooperation on papers as well as input and discussions of this thesis.

A large part of the research has been based on my involvement in cross institutional projects like the Netarchive.dk project, and the projects on overall strategy and implementation of the Danish National Bit Repository. Therefore I would also like to thank my colleagues from the State and University Library and the Danish State Archives with whom I have collaborated in these projects. Especially, I want to thank Kåre Fiedler Christensen who has been a key person in all these projects and who has also provided very useful feedback on this thesis.

Parts of the research have also been based on cooperation with colleagues from other departments of the Royal Library. In this connection I would like to thank colleagues from the Digital Infrastructure and Service Department and colleagues from the Digitisation Department. A person that I especially want to thank, for very useful and inspiring input through reviews and discussions concerning this thesis, is Birte Christensen-Dalsgaard, deputy director general at the Royal Library.

I would also like to give very special thanks to Niels H. Christensen who has given very useful input through reviews and discussions, purely motivated by interest in the subject. And special thanks to my former neighbours Eva Andersen and Mogens Larsen Andersen who have given very useful and inspirational feedback, although their PhD's are within chemistry and not computer science.

I would like to dedicate this thesis to my father Bent Zierau, who encouraged me to undertake this work shortly before his unexpected death.

Eld Zierau

Hvidovre DK, August 2011

Table of Contents

Abstract.....	6
Dansk resumé	7
List of Papers.....	8
1. Introduction	9
1.1. Digital Preservation Compared to Analogue Preservation	9
1.2. Motivation and Definition of a Holistic Approach to Bit Preservation.....	10
1.3. The Emergence of Digital Preservation as a New Research Area	13
1.4. Digital Material and Ways to Preserve Digital Material	16
1.5. The State of the Art for Relevant Areas of Digital Preservation.....	25
1.6. Contributing Results	35
2. The IR-BR Model - for Analysis of Separated Bit Preservation	38
2.1. Motivation	38
2.2. The IR-BR Model.....	39
2.3. Use of the IR-BR model.....	40
2.4. Summary.....	44
3. A Representation Concept - for Analysis of Bit Preserved Data	45
3.1. Motivation	45
3.2. The Representation Concept.....	46
3.3. Use of the Concept for Bit Preservation Representations	48
3.4. Summary.....	55
4. A Bit Preservation Evaluation Methodology - for Choice of Solution.....	56
4.1. Motivation	56
4.2. The Methodology	57
4.3. The BR-ReMS Prototype	60
4.4. Use of the evaluation methodology	68
4.5. Summary.....	72
5. A Simple Example Using the Results	73
5.1. IR-BR Ingest and Representations of Digitised material	73
5.2. Evaluation of Bit Preservation Solutions	78
6. Conclusions	81
7. Further work	84
8. Terminology and abbreviations	85
9. References.....	100
Papers.....	113
Paper A. Representation of Digital Material preserved in a Library Context	115
Paper B. Preservation of Digitised Books in a Library Context	125
Paper C. Archive Design Based on Planets Inspired Logical Object Model	135
Paper D. Cross Institutional Cooperation on a Shared Bit Repository.....	141
Paper E. Evaluation of Bit preservation Strategies.....	153
Appendices.....	163
Appendix I. Detailed Calculations using Plato.....	165
Appendix II. The BR-ReMS User interface.....	173
Appendix III. The BR-ReMS Data Model.....	201
Appendix IV. An example of a WARC files.....	202

List of Figures

Figure 1	Different analogue and digital book manifestations	9
Figure 2	Holistic approach to bit preservation within digital preservation	11
Figure 3	Simplified view of services needed for functional preservation.....	19
Figure 4	Digital preservation in an environment with dissemination	24
Figure 5	A general view of a bit repository.....	29
Figure 6	Functional entities in the OAIS reference model.....	30
Figure 7	The IR-BR model contra the traditional OAIS reference model.....	39
Figure 8	Architecture of the new Danish National Bit Repository (DK-BR)	41
Figure 9	Scenarios for use of the Danish National Bit Repository (DK-BR).....	42
Figure 10	Example of representations of a book.....	47
Figure 11	Service based persistent reference.....	52
Figure 12	Storage space factor of digitised page images per file format	54
Figure 13	The bit preservation strategy evaluation methodology	58
Figure 14	Characteristics for a bit repository with service level agreements in the general view.....	61
Figure 15	The BR-ReMS main form.....	62
Figure 16	A simplified BR-ReMS data model of entities for BR and SLA information	63
Figure 17	Forms for specification of a BR	64
Figure 18	Function for a system level result characteristic	65
Figure 19	Functions for a pillar result characteristic	66
Figure 20	The form for specification of calculations of user requirements	67
Figure 21	Entering pillar results for pillar level result characteristics	68
Figure 22	Contents of files from digitisation	73
Figure 23	Ingest in IR.....	74
Figure 24	Simple representations for preservation and dissemination	74
Figure 25	Simple representations for preservation and dissemination	75
Figure 26	Contents of derived METS files.....	76
Figure 27	Ingest in IR and BR	77
Figure 28	Requirements tree	78
Figure 29	Example of input to evaluation methodology	79
Figure 30	Transforming scheme to Plato uniform scale	168
Figure 31	The BR-ReMS main form.....	174
Figure 32	The BR-ReMS main form for requirements specification.....	175
Figure 33	The requirements form for definition of requirements	176
Figure 34	Requirements main form for specification of result characteristics	177
Figure 35	The form for definition of system level result characteristics	178
Figure 36	The form for specification of function to system level result characteristics	179
Figure 37	The form for specification of pillar level result characteristics.....	180
Figure 38	The form for function to pillar level result characteristics	181
Figure 39	The requirements form for specification of calculations of user requirements.....	182
Figure 40	The BR-ReMS main form for SLA specification	183
Figure 41	The SLA form for definition of SLAs	183
Figure 42	The SLA form for specification of a SLA	184
Figure 43	The SLA specification form for specification of SLA system characteristics	185
Figure 44	The SLA specification form for values of SLA system characteristics	185

Figure 45	The SLA specification form for specification of SLA pillar characteristics	186
Figure 46	The SLA specification form for values of SLA pillar characteristics	186
Figure 47	The BR-ReMS main form for the BR implementation	187
Figure 48	The BR-ReMS BR form for specification of BR system characteristics	188
Figure 49	The BR form for editing values of BR system characteristics	189
Figure 50	The BR form for specification of BR pillars.....	189
Figure 51	The BR form for specification of BR pillar characteristics	190
Figure 52	The BR-ReMS main form for editing values of BR pillar characteristics.....	191
Figure 53	The BR-ReMS main form for calculations.....	192
Figure 54	The calculation form for intermediate results at pillar level.....	193
Figure 55	The calculation form for results for a specific pillar	194
Figure 56	The calculation form for intermediate accumulated results at system	195
Figure 57	The calculation form for intermediate accumulated results at system	196
Figure 58	The calculation form for final user requirements results	197
Figure 59	Value type form with description of possible values for a type.....	198
Figure 60	Overview of all pillar values for one pillar characteristic and pillars in a SLA	198
Figure 61	Overview of all systems characteristics.....	199
Figure 62	Overview of all pillar characteristics	200
Figure 63	The BR-ReMS data model.....	201

Abstract

This thesis presents three main results for a holistic approach to bit preservation, where the ultimate goal is to find the optimal bit preservation strategy for specific digital material that must be digitally preserved. Digital material consists of sequences of bits, where a bit is a binary digit which can have the value 0 or 1. Bit preservation must ensure that the bits remain intact and readable in the future, but bit preservation is not concerned with how bits can be interpreted as e.g. an image. A holistic approach to bit preservation includes aspects that influence the final choice of a bit preservation strategy. This can be aspects of how the permanent access to the digital material must be ensured. It can also be aspects of how the material must be treated as part of using it. This includes aspects related to how the digital material to be bit preserved is represented, as well as requirements for confidentiality, availability, costs, additional to the requirements of ensuring bit safety. A few examples are:

- The way that digital material is represented in files and structures has an influence on whether it is possible to interpret and use the bits at a later stage. Consequentially, the way bits represent digital material influences the entire preservation of the digital material.
- The file formats can be more or less vulnerable to bit errors. Different file formats will therefore require different bit preservation solutions in order to obtain as the same maximum level of preventing risks of losing the digital material.
- The file formats can consume more or less storage volume, and the way the material is produced can influence storage volume. Thus the chosen representation and production can influence the requirements for storage volume needed in bit preservation, which influences the costs.
- There will be requirements for the availability of the bit preserved digital material in order to meet requirements on use of the digital material, e.g. libraries often need to give fast access to preserved digital material to the public, i.e. the availability of the bit preserved material must support the use.
- The digital material must be preserved in a way that satisfies general requirements for the digital material. For instance, confidentiality requirements for confidential material. Thus these requirements influence the requirements for the bit preservation.

The examples show that it is relevant to take a holistic approach and include aspects of digital representation, confidentiality, availability, bit safety and costs when defining requirements for the bit preservation. Analysis of such requirements and choice of the final bit preservation solution can be supported by the three main results presented in this thesis:

- First, a *model* to assist in analysis of a delimited and possibly shared bit repository which has to ensure preservation of bits. The model is the basis for defining terminology as well as the basis for analysis of functions within the bit repository and interface to the bit repository. This includes both technical and organisational aspects of a bit repository.
- Second, a representation *concept* to assist in analysis and design of representation of digital material under bit preservation. The concept is partly inspired by observations of possible conflicts between requirements for preservation and requirements for dissemination. The concept also includes aspects of representation assisting future use. It is furthermore inspired by results which show that decisions on the digitisation process can have impact on bit preservation costs.
- Third, a *methodology* to assist in evaluation of the best choice of a bit preservation solution, which can best meet the requirements for specific digital material. Considerations in selection of a bit preservation solution include many of the aspects from the holistic approach where various requirements for the bit preservation must be taken into account. The evaluation can be needed at different stages, e.g. in re-evaluation when a media migration is performed.

Dansk resumé

Denne afhandling præsenterer tre hovedresultater til en holistisk tilgang til bitbevaring, hvor det ultimative mål er at finde den optimale bitbevaringsstrategi for specifikke digitale materiale, som skal bevares digitalt. Digitale materialer består af sekvenser af bits, hvor en bit et binært ciffer som kan have værdien 0 eller 1. Bitbevaring skal sikre at bittene forbliver intakte og læsbare i fremtiden, men bitbevaring omhandler ikke, hvordan bittene kan fortolkes, f.eks. som et billede. En holistisk tilgang til bitbevaring omhandler aspekter som påvirker et endeligt valg af en bitbevaringsstrategi. Dette kan være aspekter af, hvordan permanent tilgang til det digitale materiale sikres. Det kan også være aspekter af hvordan materialet til bitbevaring skal behandles i forbindelse med brug af det. Dette inkluderer aspekter relateret til repræsentation af det digitale, samt hvilke krav der er til fortrolighed, tilgængelighed, omkostninger, udover kravene til sikring af bit sikkerhed. Et par eksempler er:

- Den måde, som digitalt materiale er repræsenteret i filer og strukturer, har indflydelse på, om det er muligt at fortolke og bruge bittene på et senere tidspunkt. Derfor har måden, som bittene repræsenterer det digitale materiale, indflydelse på hele bevaringen af det digitale materiale.
- Filformater kan være mere eller mindre sårbare overfor bitfejl. Derfor vil forskellige filformater kræve forskellig bitbevaringsløsning for at sikre det samme maksimale niveau for at undgå risici der fører til tab af det digitale materiale.
- Filformater kan være mere eller mindre pladskrævende, og måden, hvorpå materialet er produceret, kan have indflydelse på pladskravene. Derfor er den valgte repræsentation og produktion afgørende for krævet lagerplads, hvilket igen har indflydelse på omkostningerne ved bitbevaringen.
- Der vil være krav til tilgængeligheden af det bitbevarede materiale for at krav til brug af materialet kan opfyldes, f.eks. skal biblioteker give hurtig offentlig adgang til bevaret materiale, dvs. tilgængeligheden af det bevarede materiale skal understøtte brugen af det.
- Det digitale materiale skal bevares på en sådan måde, at generelle krav til materialet imødekommes, f.eks. fortrolighedskrav til fortroligt materiale. Derfor har disse krav indflydelse på kravene til bit bevaringen.

Eksemplerne viser, at det er relevant at tage en holistisk tilgang og inkludere aspekter af digital repræsentation, fortrolighed, tilgængelighed, bitsikkerhed og omkostninger i overvejelserne, når krav defineres for bitbevaringen. Analyse af sådanne krav samt valg af bitbevaringsløsning kan understøttes af de tre hovedresultater præsenteret i denne afhandling:

- Det første er en *model*, som understøtter analyse af et separat og muligvis fælles bitmagasin, som skal sikre bevaring af bits. Denne model er basis for definition af terminologi og basis for analyse af funktioner i bitmagasinet samt grænseflade til bitmagasinet. Dette omfatter både tekniske og organisatoriske aspekter af et bitmagasin.
- Det andet er et *repræsentationskoncept*, som understøtter analyse og design af repræsentation af digitalt materiale under bitbevaring. Konceptet er delvist inspireret af observationer af, at krav til bevaring og til tilgængeliggørelse kan være modstridende. Konceptet indeholder også aspekter omkring fremtidig brug af en repræsentation. Endvidere er det inspireret af resultater, som viser, at en digitaliseringsproces kan have betydning for omkostningerne af bitbevaring.
- Det tredje er en *metode*, som assisterer i evaluering af bedste valg af en bitbevaringsløsning, som kan understøtte kravene for et specifikt digitalt materiale. Overvejelser i forbindelse med et valg af en bitbevaringsløsning inkluderer mange af aspekterne fra den holistiske tilgang til bitbevaring, hvor der skal tages hensyn til diverse krav til bitbevaring. Evalueringen kan være påkrævet i forskellige faser, f.eks. i en reevaluering, når der skal foretages mediemigrering.

List of Papers

The present thesis is based on the five listed peer-reviewed papers. For all the papers, the peer-reviews have consisted in review by 3-4 reviewers, where the authors did not know the reviewers. For iPRES and ECDL papers, authors of the papers have been known to the reviewers, but reviewers did not know the identities of the other reviewers. The ICDL paper has both been through peer-reviews for the ICDL 2010 conference and the Journal of the World Digital Libraries.

The papers are listed in alphabetic order by the authors. Throughout the thesis, the papers will be referred to by the capital letters indicated in the list. The published papers are attached in a separate part of this thesis:

- A Zierau, E.: *Representation of Digital Material preserved in a Library Context*, In: Proceedings of the 7th International Conference on Preservation of Digital Objects, Vienna, Austria, pp. 329-337, Copyrights held by Oesterreichische computer gesellschaft, Printed by Börse Druck, www.boersedruck.at, ISBN 978-3-85403-262-5 (2010)
Full peer-reviewed paper
- B Zierau, E., Jensen, C.: *Preservation of Digitised Books in a Library Context*, In: Proceedings of the 7th International Conference on Preservation of Digital Objects, Vienna, Austria, pp. 61-69, Copyrights held by Oesterreichische computer gesellschaft, Printed by Börse Druck, www.boersedruck.at, ISBN 978-3-85403-262-5 (2010)
Full peer-reviewed paper
- C Zierau, E., Johansen, A.S.: *Archive Design Based on Planets Inspired Logical Object Model*, In: Proceedings of the 12th European Conference on Research and Advanced Technology for Digital Libraries, pp. 37-40, Christensen-Dalsgaard, B., Castelli, D., Jurik, B.A., Lippincott, J. (eds.) LNCS, vol. 5173, Publisher: Springer-Verlag Berlin Heidelberg, ISBN 978-3-540-87598 (2008)
Short peer-reviewed paper
- D Zierau, E., Kejser, U.B.: *Cross Institutional Cooperation on a Shared Bit Repository*, In: Journal of the World Digital Libraries, vol. 3, issue 1, pp. 11-21, Publisher: TERI Press, New Delhi, ISSN 0974-567-X (2010)
(Awarded with *best paper award – international* at *International Conference on Digital Libraries*, New Delhi, India, 2010)
Full peer-reviewed paper
- E Zierau, E., Kejser, U.B., Kulovits, H.: *Evaluation of Bit Preservation Strategies*, In: Proceedings of the 7th International Conference on Preservation of Digital Objects, Vienna, Austria, pp. 161-169, Copyrights held by Oesterreichische computer gesellschaft, Printed by Börse Druck, www.boersedruck.at, ISBN 978-3-85403-262-5 (2010)
Full peer-reviewed paper

1. Introduction

The subject of this thesis is within the area of digital preservation. *Digital preservation* is defined as the series of managed activities necessary to ensure continued access to digital information for as long as necessary [31,71,72]. Digital preservation is also referred to as permanent access [55].

Throughout the chapters, terms that are defined will be highlighted by *italic and underscore*, which also will be the case for the names of the three main results, if relevant.

1.1. Digital Preservation Compared to Analogue Preservation

Digital materials differ from analogue materials in various ways. This includes the media on which they are kept, the mechanisms to interpret the digital information from the media in an understandable way, and the complexities in the materials. The term *digital material* is here used instead of digital information to indicate that the focus is on conceptual material in digital form, while *digital information* is more general denotation of information consisting of sequences of bits (also called *bit-streams*), where a *bit* is a binary digit which can have value 0 or 1.

A simple example of differences between digital and analogue material is different manifestations of a book. A *manifestation* is “the physical embodiment of an expression of a work” [141], e.g. manifestations of a book can be a recitation of the book, or a theatre play created on basis of the book. In the example in Figure 1 the manifestations are in form of a physical book, a recitation of the book, and a digital manifestation of the text and sound from the analogue manifestations.

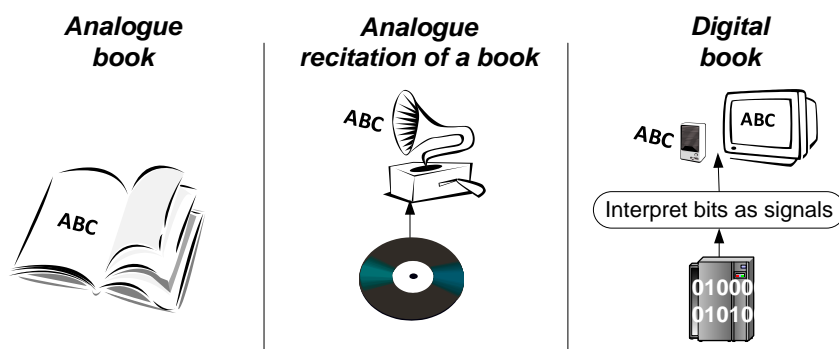


Figure 1 Different analogue and digital book manifestations

The preservation of an analogue book printed on paper consists of preservation of the physical material as well as registration of relevant *metadata*, which is structured data about data [104], for instance title, where the book is placed etc.

The preservation of the recitation of the book on the long-playing phonograph record (LP record) is more complex. The preservation will be similar to the book regarding preservation of the media, which in this case is the vinyl LP record, and metadata. The metadata must in this case also include a description of how signals can be read from the vinyl and interpreted as sound in speakers.

Preservation of the digital book manifestation must include activities to preserve the basic digital information (in the form of bits) on a digital media, the relevant metadata of the material including file formats and structures within the digital material, description of how to interpret the bits, and ensuring actual replay of digital information on various devices, e.g. speakers.

Assuming basic reading and language skills, the preserved analogue book can be read directly in the long term. The replay of the recitation of the book from an LP will require a device to read and play; a standardised sequential interpretation of information from the vinyl. The digital book also needs applications to interpret the stored bits as signals for screen or speaker, and it needs the devices to replay these interpreted signals. However, the interpretation includes complex combinations of the ability to read from the media, the collection of the necessary information to combine the parts and calculation of how the individual parts are contributing to the final rendering of the digital material.

In general, media used for storing digital information tend to be much more short-lived than analogue media. The media is the basis for accessing the original information stored on a physical media; therefore the media is essential for all preservation actions, analogue as well as digital. For many analogue materials replication is one way of preserving the material (e.g. for books) [177]. This method is also a recognised method for digital material, and with less risk than for most analogue materials. Analogue material can be authentic, e.g. as the original letter written on paper or the original printing of a book. Digital material consists of bits written on digital media, but they are not bound to the digital media in the same way. Therefore it does not make sense to talk about an authentic bit or bit-stream on a media, since the information replicated to another media is just as authentic as the one it was replicated from [177].

In digital preservation, a distinction can be made between preservation of the bits called bit preservation, and the functional interpretation of the bits called functional preservation:

- *Bit preservation* is defined as the required activities to ensure that the bit-streams remain intact and readable [86]¹, i.e. the bits on the disk illustrated in Figure 1 are intact and can be read.
- *Functional preservation*² must ensure that the bits remain understandable and usable according to the purpose of preservation, i.e. in the example in Figure 1; functional preservation must ensure that the interpretation can take place, and result in signals to devices which represent the digital material as prescribed by the preservation purpose.

1.2. Motivation and Definition of a Holistic Approach to Bit Preservation

Bit preservation is the basic part of all digital preservation activities [188]. If the bits are lost or damaged, it can lead to loss of the digital information, which disables any form of digital preservation. In particular, bit errors in compressed files or file formats that are not robust to bit errors, can mean loss of the entire contents of the files [54,192].

At a first glance bit preservation may appear as a simple challenge [53,82,188]. However, the challenge gets more complicated when there are other requirements than bit safety requirements to be considered. The term *bit safety* denotes how safe the bit-streams are from alterations or becoming unreadable. Bit safety is related to the *bit integrity* which only denotes the consistency of the bit-streams, i.e. that they stay unchanged.

The ultimate goal with bit preservation is to find the optimal bit preservation strategy that can support the various requirements. A *strategy* refers to a plan of actions designed to achieve a particular goal. The contributions presented in this thesis support such a goal of achieving an optimal bit preservation

¹ The paper “Thirteen Ways of Looking at...Digital Preservation” defines it as “... an assurance that the bit streams constituting the digital objects remain intact and recoverable over the long-term.” [86], which basically means the same.

² Functional preservations are also sometimes denoted as logical preservation.

strategy. These contributions concern: a model to support definition of a bit repository, a concept to assist in analysis and specification of various bit preservation requirements, and a methodology to assist in evaluation of which bit preservation solution that best meets requirements to be met in an optimal bit preservation strategy.

The ideas for these contributions are primarily based on work which I have carried out at The Royal Library. The work has included development of strategies for a national bit preservation solution, and requirements for the different types of digital material to be bit preserved, as well as analysis of a system to support both dissemination and preservation. From this work it has been clear that bit preservation is more than ensuring unchanged bit sequences and readability, and that a more holistic approach must be taken.

Holistic comes from holism, which is a way to contemplate elements for instance within biological and sociological sciences. It is a way to explain a phenomenon in its entirety. As an opposite of atomism, holism explains less composed phenomena by the composed. In other words, the whole cannot be expressed as a function of its parts. Instead holism emphasises the importance of the whole and the interdependence of its parts, i.e. the way parts work is seen as a consequence of the structure of the whole [50]³.

My definition of a *holistic approach to bit preservation* is an approach where bit preservation is viewed as something that must be regarded as part of a whole, where different circumstances can influence how bit preservation must be performed. It is called a holistic *approach* since this thesis does not try to make a map of all the phenomena that can affect the choice of a bit preservation solution.

The purpose with a holistic approach is to set up a framework for the aspects to be considered when seeking to find an optimal bit preservation strategy that can support various requirements coming from the “whole”. A graphic illustration of how a holistic approach covers parts of digital preservation is given in Figure 2:

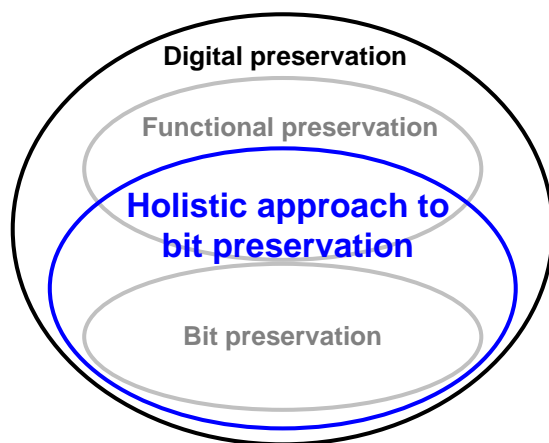


Figure 2 *Holistic approach to bit preservation within digital preservation*

The definitions of bit preservation and functional preservation include a split between the two. As illustrated in Figure 2, the holistic approach to bit preservation includes all aspects of bit preservation. The approach only includes parts of the functional preservation aspects: These aspects are the ones that

³ translated from Danish and given in revised form.

influence a choice of ways to ensure bit preservation. Note that the size of each oval in Figure 2 is of no importance to the meaning of Figure 2.

There is more to digital preservation than just bit preservation and functional preservation, for instance sustainability aspects. This is illustrated in Figure 2 by having bit and functional preservation as part of digital preservation. The holistic approach also includes some of these other aspects which may not be classifiable as bit preservation or functional preservation aspects. Thus the holistic approach is illustrated to include aspects of digital preservation that are neither part of bit preservation nor part of functional preservation.

It could seem that a holistic approach to bit preservation could be just another way of describing digital preservation. This is however not the case. The approach only contains the aspects that influence bit preservation. For example, detailed theory on functional preservation actions is not part of the holistic approach, although there may be aspects related to actually accessing and rendering bit preserved material which must be covered in the holistic approach to bit preservation.

Bit preservation must at some point be ensured by a bit preservation solution. The holistic approach focuses on the requirements for such a bit preservation solution, when the bit preserved material and bit preservation solution are viewed as part of the whole.

Requirements for a bit preservation solution can also be expressed as requirements for a system with a surrounding organisation which represents a bit preservation solution. This will be denoted as a bit repository. A *repository* is for analogue materials defined as a building or room designed or arranged and used specifically and exclusively for long-term storage of archive and library materials [64]. For digital material this is, however, not bound to buildings or rooms. It is instead the organisation and techniques designed and arranged and used specifically and exclusively for long-term storage of digital materials. Using this definition of a repository, requirements for a bit preservation solution can also be expressed as requirements for a bit repository.

There are various examples of requirements coming from “the whole” which influence requirements for and choice of a bit preservation solution. These are examples of a broader context of bit preservation covered by a holistic approach to bit preservation, which is not only concerned with bit integrity and readability of bits.

Integrity is one of the information security aspects defined in the ISO 27000 series [68]. *Information security* aspects cover:

- *integrity* which is the property of safeguarding the accuracy and completeness of digital material (e.g. bit integrity for bit in a bit repository)
- *availability* which is the property of being accessible and usable upon allowed demand (e.g. ability to access and possibly process the bit-streams)
- *confidentiality* which is the property that information is not made available or disclosed to unauthorised individuals or processes (e.g. only a restricted group of users are allowed to access the bit-streams)

The ISO 27000 standard does, however, not directly address bit safety concerns as part of integrity, since it is aimed at repository in general, where bit preservation is just a part of the repository’s functions. All the information security aspects are relevant for requirements for bit preservation. Another important example of additional requirements is cost requirements (e.g. respect of budgets for preservation).

The holistic approach includes many aspects of the representation of bit preserved digital material, and there are many examples of requirements which relate to various aspects of the bit preserved material. However, before this can be explained in more detail, a more thorough introduction of digital material and ways to preserve it is needed. This introduction is given in section 1.4, which will include examples of which aspects belong to a holistic approach, and which aspects do not. Before this introduction, there will be a description of the background for interest in digital preservation as a new research area.

1.3. The Emergence of Digital Preservation as a New Research Area

Digital preservation is a relatively new area. It was not until the mid-1990s that the challenges and urgent need for digital preservation was clearly articulated [47]. At that point much digital information had already been lost [95,147].

1.3.1. The emergence of the concept of digital preservation

In the early phase of digital preservation there was attention on the fragility of digital media. In many cases, the preservation strategy for digital material was to preserve them in analogue form by printing them from less stable magnetic and optical media to papers and microfilms [53], but there was also awareness that not all digital material could be preserved this way [53, 82,188].

Digital preservation has emerged along with technological evolution. During the last two decades, digital material has gained an increasing importance for many business areas within private industries and public services. One example is the pharmaceutical industry, where legislation requires documentation of development and production of medical products in order to ensure the safety of the products to be marketed. Specifically, pharmaceutical products sold in USA must meet regulations specified by the U.S. Food and Drug Administration (FDA) [180], which has strong requirements for the record keeping of both development and production of drugs [127]. Another sector is that of archives and libraries, which must comply with legislation, e.g. legal deposit laws [129]. The digital material that must be preserved is increasingly material created in digital form, e.g. e-mails from authors and researchers [73], the national domains of the World Wide Web [63] and e-books [59], and digital observation data in space agencies [11].

Particularly for archives and libraries, the challenge was to ensure the digital heritage for the future. The main challenge was to establish trustworthy and sustainable repositories that could take the responsibility of preserving digital material. Part of this challenge relates to the nature of the digital material, ensuring funding and legal issues, establishment of technical infrastructure, standards and best practices as well as means to communicate these on national and international levels [53,82,188].

Ensuring sustainable digital preservation involves a wide range of issues that must be dealt with, not just bit preservation and functional preservation. This is expressed in “Thirteen Ways of Looking at...Digital Preservation” in the following way [86]:

“Preserving our digital heritage is more than just a technical process of perpetuating digital signals over long periods of time. It is also a social and cultural process, in the sense of selecting what materials should be preserved, and in what form; it is an economic process, in the sense of matching limited means with ambitious objectives; it is a legal process, in the sense of defining what rights and privileges are needed to support maintenance of a permanent scholarly and cultural record. It is a question of responsibilities and incentives, and of articulating and organizing new forms of curatorial practice. And perhaps most importantly, it is an ongoing, long-term commitment, often shared, and cooperatively met, by many stakeholders.”

These issues were also addressed by the US task force who in 1996 published their report on archiving of digital information, which identified the need for a deep infrastructure for digital archiving and long-term preservation [188]. The purpose of the task force was to investigate the means of ensuring continued access to digital information indefinitely. This report has frequently been referred to [53,82,146,177], and according to the “JISC Beginner’s Guide to Digital Preservation” [175], the report has had impact world-wide. In short, the conclusion of this report was that a range of initiatives was needed in order to secure funding and support evolution of trusted preservation services by means of national and international cooperation, technical infrastructure, practical experience, best practices and standards.

In 2002, a short status was presented in the paper “Introduction: The Changing Preservation Landscape” [98] on some of the significant work that had been done since the report of the task force in 1996. This status included mention of the establishment of several national organisations to support digital preservation initiatives, for instance Digital Preservation Coalition (DPC) in the United Kingdom, establishment of an Australian subject gateway to digital preservation resources called Preserving Access to Digital Information (PADI), and work on a US national strategy for preserving digital information under development by the Library of Congress, which became The National Digital Information Infrastructure & Preservation Program (NDIIPP). Furthermore, the status mentioned work on establishment of best practices by Online Computer Library Center (OCLC) and Research Libraries Group (RLG).

1.3.2. Recognition of digital preservation as a research area

Also in 2002, a working group was established by the United States National Science Foundation’s (NSF) and European Commission sponsored Network for Digital Libraries under the Fifth EU Framework Programme (DELLOS) [28,114]. This working group was to address the challenges surrounding long-term digital preservation and curation [146]. Although a wide range of initiatives and organisations had been established since the report from the Task Force was published in 1996, this working group reported in 2004 on a wide range of issues that still needed to be addressed in research. It recognised that robust, effective, and affordable strategies for digital preservation depend on an infrastructure of common standards, methods, and tools, but also that an infrastructure for digital archiving requires more than research and development of tools and technologies.

The outcome of the working group from 2002 has set digital preservation on the agenda as a research area within the EU research programmes. Since then, mainly starting in the 6th and 7th programme, there have been various research focuses within digital preservation as part of the EU research programmes. The objectives for the research have moved from a library/archive centric view to one that is increasingly focused on understanding the challenges posed by the nature of the digital content itself, and now also focused on organisations that are only beginning to face the problems of preserving their digital material [97]. One example of an EU research projects is the Planets (Preservation and Long-term Access through NETworked Services) project which was to build practical services and tools to help ensure long-term access to digital material. The Planets project has also contributed to some of the research presented in the papers on which this thesis is based. Another example is the SCAPE (SCALable Preservation Environments) project which among other aspects investigates scalable preservation actions. The SCAPE project will contribute to the holistic approach to bit preservation with aspects of mass processing on bit preserved material. *Mass processing* here means processing of large data archives, analogously to mass digitisation [155].

There are many other research projects established both in the EU and the rest of the world. The research areas within digital preservation are many. The above examples are mainly focused on infrastructure.

Other areas are for instance costs aspects, auditing and work flow provenance. Bit preservation has, however, mainly been left to the hardware industry, which will be explained later in this thesis.

1.3.3. Minimal focus on bit preservation

Especially in the literature from the 1990's, there are many examples of loss of digital materials [95,147]. Many of these losses are caused by fragility and lack of readability of digital media, for instance from The BBC Domesday Project from 1986 which resulted in a large volume of digital information, where a large investment was made 14 years later to try and rescue the data from obsolescence or destruction due to media faults [14]. An example of losses of data due to deteriorating or damaged storage media can also be found in the NASA Viking Mars Mission, where 10-20% of the data was lost [145]. Losses can also be due to whether bits can be read from the media. This was for example close to happening to a collection of DAT-tapes at the State and University Library in Denmark, where the contents of the tapes needed a specially developed media migration tool in order to be media migrated into a readable form [194]. As discussed later there can be many more reasons for loss. However, the loss of data has been a sensitive topic, which few will admit to have suffered [95]. There is still loss of data [192] and probably not all the cases have yet been made public.

In the late 1990s the actual handling of the physical storage was seen as the least of the worries [82]. As expressed in the report "Preserving Digital Information: Report of the Task Force on Archiving of Digital Information" [188] p. 12 (with reference to [48,93,94]):

"There are various well-established techniques, such as checksums and digests, for tracking the bit-level equivalence of digital objects and ensuring that a preserved object is identical to the original".

Here a *digital object* is defined as an object composed of a set of bit sequences [16], i.e. digital material is represented by digital objects. Tracking equivalence between identical digital objects is one of the easy tasks for digital material unlike most analogue materials, since digital objects can be replicated at any time to different media without changing the digital object itself. In the rest of this thesis a *replica* will be used for a copy of the data stored in a technical environment within an organisation as the result of a replication. Replication as part of exchanging the media because of ageing is called *media migration*.

The techniques to replicate and check the basic digital data in form of bits are the basis for bit preservation. The initial impression of bit preservation, as being an issue as good as solved, has meant that it has had limited focus for some time. An example is that the paper "The State of the Art and Practice in Digital Preservation" [88] refers to the task force report "Preserving Digital Information: Report of the Task Force on Archiving of Digital Information" [188] as evidence that bit preservation is out of focus. Another reason that bit preservation has been out of focus is that there have been great technical enhancements of storage technology over the last decades. However, as discussed later, bit preservation is not as solved as it seems [143], and there are still many unanswered questions of how to choose an optimal bit preservation solution for specific digital material.

1.3.4. Practical developments and knowledge exchange

All over the world projects have continuously been initiated to construct practical solutions, and various libraries have developed, and host, e.g. work on standards, registries and best practices. There will be more about relevant projects and contributions from libraries and archives in section 1.5 "The State of the Art for Relevant Areas of Digital Preservation" after the introduction of digital material and digital preservation strategies in section 1.4.

Knowledge exchange on digital preservation is made through many different channels, for example, through different knowledge exchange programs (e.g. Knowledge Exchange [78]). Digital Preservation has also started to become part of the research institutions for computer science, for example the Digital Preservation group at Vienna University of Technology [185].

There have also emerged a number of conferences and workshops dedicated to preservation issues, for example International Conference on Preservation of Digital Objects (iPRES) and the International Digital Curation Conference (IDCC). Furthermore a wide range of conferences concerned with digital libraries include digital preservation aspects, for example European Conference on Research and Advanced Technology for Digital Libraries (ECDL)⁴, International Conference on Digital Libraries (ICDL), and Joint Conference on Digital Libraries (JC DL). Lastly, there are examples of conferences and workshops that are concerned with more specialised areas of the digital library: The International Web Archiving Workshop (IWAW), and Very Large Digital Libraries (VLDL). These are followed by various journals for scientific papers which include digital preservation topics, e.g. The International Journal of Digital Curation.

Knowledge exchanged is also to a larger extent made via communities built around projects or specific solutions. There will be more about this in section 1.5 “The State of the Art for Relevant Areas of Digital Preservation”, but first there will be a section introducing the different digital materials and the ways to preserve them according to their preservation purpose and use.

1.4. Digital Material and Ways to Preserve Digital Material

Before digital material can be bit preserved, there are a wide range of issues that must be considered. This includes the source and complexity of the digital material, the preservation strategy for the digital material, and how the digital material must be accessible on a permanent basis. This includes how digital material may both need support preservation actions and actions needed for access to the material.

1.4.1. Sources and complexities

As recognised from the early days of digital preservation, there are different sources of digital information due to how the information originates, and there are varying complexities in digital materials [53,147]. Both sources and complexities influence which digital preservation actions that are required in order to fulfil the intention and purpose of the digital preservation including bit preservation. The three main groups of sources *Digitised material*, *Digital substitution of analogue material*, and *Digitally born materials* are explained here. Complexities are most common among digitally born materials, and are therefore explained together with this type of materials. Along with the descriptions there are examples of how the source and complexities can influence considerations on how the digital material must be digitally preserved.

Digitised material is analogue material transcribed to digital material which can represent the analogue material in a digital form. An example of digitised material is a project at the Royal Library where the Danish national literature from before year 1700 is digitised [119]. There can be different purposes for the digitisation, and thus different purpose and intention with preservation of the digitisation. For example, in some cases the purpose of a digitisation can be to make the analogue material more accessible. The preservation could in this case be to protect the investment in the digitisation.

⁴ from 2011 this conference has been renamed to TPD L which is an acronym for International Conference on Theory and Practice of Digital Libraries.

Digital substitution of analogue material is a special case of digitisation, where the purpose of the digitisation is to create a digital copy of a non-digital material in order to support preservation [75]. In this thesis the term is used in cases where the substitution copy substitutes the original as the master for further preservation. An example of digital substitution is the deteriorated negatives at the Royal Library of Denmark which will be digitised and digitally preserved [76]. Here the choice of using digital substitutions rather than analogue substitutions is based on a study of a cost-benefit analysis for the different preservation possibilities [75]. With this definition of digital substitution, the digital material is the new preservation master substituting the analogue material, and thus from a preservation perspective the digital material must be treated as digitally born material.

Digitally born materials are digital materials that are created in digital form. Examples of digital materials that are born in digital form are images from digital cameras, emails and blogs on the World Wide Web. There are many examples of digitally born materials that have been recognised as complex, e.g. databases, geographic information systems, virtual reality models, and the World Wide Web [53,147,177].

In many cases, it does not make sense to make an analogue preservation of digitally born materials. Already in the early days of digital preservation, it was recognised that the World Wide Web would be hard to preserve in an analogue form, because of the linkage of material, and the variety of file format [53,147]. Since then, the challenge has grown bigger as there are increasing number of formats, increased complexity in formats, changed technologies to operate on formats, and emerging interactive programs [92]. Another factor is whether it is feasible to preserve it in analogue form. A concrete example is the archive of the Danish domains of the World Wide Web [116]. Web data can include any kind of digital material e.g. images, sound, videos, and it can contain links embedded in the text. Thus, if the purpose is to preserve the nature of the web following links, playing videos etc. then the only alternative is to preserve it digitally. Another reason to preserve it digitally is that the volume of the present web archive of the Danish web of 100 Terabytes is roughly estimated to be a 100 km high pile of printed pages⁵ in analogue form.

Another example of a complex digitally born material is computer games. Computer games have also become more advanced during the last decades. Today, it is not only a problem that many computer games must be executed on special devices; one of the big challenges is social computer games, where a lot of the activity goes on as a social activity on the World Wide Web [99], for example in the game World of Warcraft [99,191].

A third example of a complex digitally born material is scientific data. Scientific data both covers e-science and actual complex data from research. E-science is a concept that has emerged from the more and more advanced use of digital information in research communities, where researchers want to preserve and share their science across institutions and borders [2]. For example, the space industries are collecting still more digital information from observations that preferably should be comparable across continents [11]. The challenges are many, as also is described in the book "Advanced Digital Preservation" [44].

The complexity in digital materials means that it becomes even more crucial that all relevant information is available on a long-term basis in order to make the digital material accessible in the future. This includes being able to identify, access and understand the future versions of the material, where all relevant metadata is ensured to be available.

⁵ The background for this measure is provided in the terminology chapter under description of 'netarchive.dk'.

1.4.2. Preservation actions and preservation strategies

The preservation actions in bit preservation are mainly related to ensuring bit integrity and making timely media migration. The actions do not change the bit-sequences. Thus the form of the digital material is not influenced by bit preservation actions after it is sent to bit preservation⁶. In this description of digital material there will therefore not be further description of preservation actions for bit preservation. More description will instead be provided on the state of the art of bit preservation in section 1.5.

Turning the focus to functional preservation, there are cases of functional preservation actions that can influence the form of digital material, which will be described in more detail below. The challenge of functional preservation is one of the major concerns addressed in the reports and papers from the 1990's [53, 82,188]. There are many reasons why digital information becomes obsolete, and thus continuously provide challenges to digital preservation. Some of the main reasons are rapid changes in means of recoding, ever increasing variety of file formats and integrated document functionalities, and running on different platforms [47,136,177]. There are therefore many aspects of functional preservation which have been the target for research during the last two decades.

This thesis does not intend to give a full picture of all aspects of functional preservation, but only an overview of some of the aspects that are relevant for the holistic approach to bit preservation. This will include functional preservation strategies and a short overview of preservation actions in functional preservation.

Functional preservation strategies

There are three main preservation strategies for digital preservation, which have been recognised as main preservation strategies already from the mid 1990's [188]. These preservation strategies are [71,88,136]:

- Technology preservation (also called technological museum)
The Technology preservation strategy consists of preservation of the needed computer hardware and software platforms which the digital material can be rendered on [92,188].
This technology is rarely used, since it will be expensive, if not impossible, to maintain all hardware and software platforms on a long-term basis.
- Emulation
The emulation strategy consists of simulating the original environment that was used to render the digital material. The original bit-streams are then rendered in a new environment via the emulated environment [74,88,181].
There are different interpretations and set-ups for emulation. However, it is out of the scope for this thesis to go into further details. There are various projects that have worked with emulation. Some examples are the KEEP⁷ (Keeping Emulation Environments Portable) project [74] and Preserving Virtual Worlds project [99].
In spite of the work that has been done, there is still a lot of work to be done before the emulation strategy can be fully supported [74,99].
- Migration
The migration strategy consists of migration of the data from one representation to another, i.e.

⁶ There will be changes in form of updates of audit trails telling that an action has been taken, but these can be treated separately.

⁷ the KEEP project is another example of a EU research project.

from one structure and contents represented in a set of files to a possibly new structure and a new set of files with new file formats [88,194].

The migration strategy is the preservation strategy with which there is most experience, both in practice for simple cases like migration into formats with different character encoding [177], and in evaluation of whether a specific migration from one format to another was suitable for specific materials (e.g. see [46,81]).

Migration is however also a preservation strategy where the bit streams are changed entirely, and thus it is harder to argue for the authenticity of the digital material after a migration [45]. According to the new draft recommended update of the Open Archival Information System (OAIS) reference model⁸ [17], *authenticity* is defined as the degree to which a person (or system) may regard an object as what it is purported to be [16,17,45].

The most evolved and used preservation strategies are emulation or migration strategies [6]. For both of these preservation strategies the success of preservation depends on the success of preserving the authenticity in an emulated environment or in the target file format of a migration.

The holistic approach to bit preservation includes consideration of how the different preservation strategies can be supported, e.g. by enabling some sort of access to metadata needed for a preservation strategy. Such considerations influence the form that the digital material must have when bit preserved. However, theoretical details on how to do emulation do not form part of the holistic approach, if no requirements are consequently set to the bit preserved material.

Managed activities in functional preservation

Functional preservation covers the interpretation of bits. Thus functional preservation, as part of digital preservation, will consist of the series of managed activities that ensures that the bits remain usable and understandable for their defined purposes.

A rough sketch of categories of actions needed in functional preservation is illustrated in Figure 3. This is simplified version of an overview which can be found in the Planets project⁹. In this sketch the actions are expressed as services needed for functional preservation.

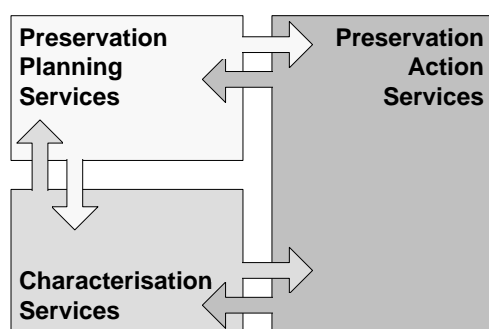


Figure 3 Simplified view of services needed for functional preservation

The arrows in Figure 3 illustrates that the services are interlinked, which will be explained in the description of the different services:

⁸ More details on OAIS will be given later in this thesis.

⁹ See e.g. the Planets overview presentation http://www.planets-project.eu/docs/presentations/Planets_overview_2006_06.ppt which can be found via the Planets website [128].

- Characterisation

Characterisation consists of finding characteristics of digital material and file formats. Characteristics are property/value pairs for an aspect, in this case describing aspects of the file formats. That means characterisation involves extraction of information that describes digital objects, e.g. metadata, file format features and aspects of content [7]. An example of a format characterisation tool is JHOVE [1]. The characterisation is needed for the following:

- To validate detailed functional preservation strategy. For example whether an explicit migration tool can be used for specific digital materials where specific properties are required to survive. The XCL language is an example of a language that can support such characterisations and evaluation [7,172].
- As input to preservation planning which includes technology watch. Technology watch covers tracking emerging digital technologies, information standards and computing platforms (i.e. hardware and software) to identify technologies which could cause obsolescence in the archive's computing environment and prevent access to some of the archives current holdings. An example is technology watch of file formats, used in the repository, and is endangered of becoming obsolete.
- As input to consideration of whether a specific file format is suitable as a preservation format. Preservation formats are formats that are accepted for long-term preservation. In order to be able to do functional preservation, preservation formats are for instance usually required to be standardised formats.
- As part of validation of whether a file conforms to file format specification of the chosen preservation formats. Such validation is usually required before the digital material is archived.

The characterisation and related activities usually involves format registry containing various information about different file formats. PRONOM is an example of such a registry [176].

Characterisation is an example of an aspect that may be considered in a holistic approach to bit preservation. Characterisation information can e.g. be available as part of the metadata for bit preserved material, or by enabling the possibility of doing mass processing with characterisation tools on the bit preserved material. However, there are also aspects of characterisation that are not relevant for a holistic approach to bit preservation, which is the case for details in theories behind the XCL language.

- Actions

Actions cover different preservation actions for preservation strategies performed by tools and procedures. This includes tools to support emulation and migrations. Further examples of such tools can also be found in the paper on “The Planets Approach to Migration Tools” [194] and in the KEEP project [74].

The choice of tools for preservation action depends on the ability of the tools to give an authentic version of the digital material as a result of the migration action.

As noted in the section on preservation strategies the holistic approach includes aspects of enabling the different migration strategies on bit preserved material.

- Planning

Planning involves specification of preservation plans as well as determining the best detailed preservation strategy [6]. Preservation actions are initiated based on the preservation plans, which again are based on e.g. tools information, characterisation information and technology watch.

There are pre-conditions for planning and timely execution of appropriate functional preservation actions, for example that there is sufficient information, or access to retrieval of information, on which the planning is based. That means that the bit preserved digital material must be prepared for planning and execution of functional preservation actions.

Common to all the above categories of actions is that they are either based on or contribute to the metadata of the digital material.

1.4.3. Metadata

Metadata contains records of the context of the digital material. Preserving the context of digital materials can be just as important as preserving the actual data [4,5]. For example, if the bit preserved digital material includes a TIFF 6.0 file then part of the preserved information needs to contain this information, in order to enable rendering or enable functional preservation actions. The importance of metadata is expressed in the document “Metadata for digital libraries: state of the art and future directions” [43] p. 5 as follows:

“Metadata is the core of any information retrieval system and so its implications for any digital library are profound: the choice of a metadata scheme underpins any such library's ability to deliver objects in a meaningful way, and greatly affects its long-term ability to maintain and preserve its digital assets.”

This will apply for any institution or company that has obligations to preserve and to give access to digital material.

It is commonly accepted that there are different types of metadata, for example technical metadata is needed in order to plan and perform digital preservation. This thesis will not attempt to give a full list of different types of metadata, but it will here give an example of some main types of metadata [43] as well as examples of metadata schemes that can contain such metadata:

- Descriptive metadata
The information describing the intellectual contents of digital objects. Examples of descriptive metadata are author, title, and publisher.
An example of a standard schema for descriptive metadata is MODS (Metadata Object Description Schema) [109].
- Administrative metadata
The information necessary to curate the digital object. This usually includes:
 - Technical metadata
The technical information, for example file formats or how an image was scanned.
There are different standard schemas for details of different types of digital files. Examples are MIX (Metadata for Images in XML) for still images [108] and TEI (Text Encoding Initiative) for texts [169].
 - Rights management
The information necessary to restrict its delivery to those entitled to access it.
There are different standard schemas that can specify these metadata. Examples are METS (Metadata Encoding and Transmission Standard) [102] and PREMIS (Preservation Metadata: Implementation Strategies) [132].
 - Digital provenance
The information on the creation and subsequent treatment of the digital object, including details of responsibilities for each event in its lifespan.

This is for example the place where metadata describing preservation actions must be placed. The best known schema for these metadata is PREMIS.

- Structural metadata

The information on the internal structure of a digital object, so that it can be rendered. An example is the page order for a book, if pages are given in separate files.

The METS standard and schema includes such metadata.

Especially METS and PREMIS have played important roles in creation of best practices within digital preservation, therefore more information on these standards will be found in section 1.5 “The State of the Art for Relevant Areas of Digital Preservation”.

As the metadata describes the context of a digital material, these metadata can be an important part of the data that must be bit preserved. Metadata is thus important in consideration of the representation of digital material put to bit preservation in a holistic approach to bit preservation. There will therefore be more on metadata in chapter 3 “A Representation Concept - for Analysis of Bit Preserved Data”.

1.4.4. Identification and versions of digital material through preservation

A very basic precondition, of making digital material accessible in the future, is that the material can be identified. Identification of digital material is one of the extra challenges compared to analogue, since the digital material can change as consequence of the digital preservation actions or added information [13,38].

If a version of digital material is a result of a preservation action, then the new version must keep digital material understandable and usable for its defined purposes. This is also referred to as preservation of the authenticity of a digital object, where the degree of authenticity is judged on the basis of evidence [17]. The evidence will for instance consist of provenance and audit trail information which includes all history of access of the object, changes in object, or changes for the object.

Preserving authenticity can also be expressed as preservation of the significant properties of the digital material. A widely accepted definition of significant properties according to the paper “Significance Is in the Eye of the Stakeholder” [26] p. 298 (referring to [190]) is:

“The characteristics of the digital objects that must be preserved over time in order to ensure the continued accessibility, usability, and meaning of the objects, and their capacity to be accepted as evidence of what they purport to record”¹⁰.

However, it can be questioned how significant properties can be chosen in order to preserve the authenticity of the digital material [45]. As stated in the paper “Excuse me...some digital preservation fallacies?” [150]:

“The problem here is that there is no way of precisely defining the designated community, and similarly no way of foretelling the properties that future users might deem significant. This leads to pressure for preservation that must be faithful to the original in all respects”.

Preservation that is faithful to the original in all respects may not be possible. Even with a limited set of significant properties, it may not be possible to get a new version which contains all the required significant properties. An example could be a PowerPoint file with a presentation that includes

¹⁰ There is not complete consensus on how significant properties should be defined [45]. This differences is however not important for this thesis.

animations, sound tracks, links to the World Wide Web, and Excel sheets with embedded calculations. In case the specific PowerPoint format becomes obsolete, several migrations into file formats which include different significant properties may be needed, for example a PDF file with images of the slides which can display the original fonts, shapes, and colours, and a file as a result of migrating to a newer version of PowerPoint where the different layers in animation can be identified.

In the example of the PowerPoint file, the sequence of migrated versions of the file is not linear, since the old version will be branched, and where each branch can have its own version history. This gives a challenge for identification of the digital material over time. However, there are many cases where version sequence is assumed to be linear, and which therefore do not address this challenge [13,25,38,160,171].

Persistent identifiers are identifiers of the digital material that are persistent over time, i.e. they cannot be reused for other digital object. However, it is not fully clear what the persistence consists of. Ideally the persistence is both concerned with the identifier itself and the authenticity of the object it points to. However, such a definition of persistent identifiers is not just a technical issue. As stated in “Advanced Digital Preservation” [44] p.179:

“To produce general purpose Persistent Identifiers, which could be used to point to any or all objects, is well known and challenging, the difficulty being social rather than technological”.

The reason is that definition of what a persistent identifier points to will depend on definition of what is defined to be the authentic version of the digital object. The fact that an object cannot be assumed to have a linear version history, complicates this definition even more. Other challenges to definition of persistent identifiers are how they can be resolved in the future. The paper “Persistent Identifiers: Considering the Options” describes it as follows [178]:

“Persistent identifiers (PIs) are simply maintainable identifiers that allow us to refer to a digital object – a file or set of files, such as an e-print (article, paper or report), an image or an installation file for a piece of software. The only interesting persistent identifiers are also persistently actionable (that is, you can “click” them); however, unlike a simple hyperlink, persistent identifiers are supposed to continue to provide access to the resource, even when it moves to other servers or even to other organisations. A digital object may be moved, removed or renamed for many reasons.”

This description points at various challenges for persistent identifiers as a resource, but it also has an implicit assumption of linear version history.

As identification of the digital material is crucial for the entire digital preservation, there will be more on this in chapter 3 “A Representation Concept - for Analysis of Bit Preserved Data”.

1.4.5. Different actions needed on digital materials

In most cases digital material must be prepared for use as well as being preserved. For instance, most public library material must be accessible at least on a day-to-day basis or even within seconds via services offered via the World Wide Web. In such a case, this will mean that the actions needed for the use of the material may set requirements for accessibility of the bit preserved digital material. In many cases the additional requirements will relate to dissemination of the digitally preserved material. This thesis will refer to *dissemination* as the process of providing information to an access point for a user of the repository¹¹.

¹¹ Dissemination is more precisely defined in OAIS terms in the terminology chapter.

Figure 4 illustrates digital material as part of a system where the digital material must both preserved and disseminated. The illustrated intersection between preservation and dissemination has some sort of joint model of representations for the two (the dashed circles illustrate which parts of the joint model correspond to representations in preservation and dissemination respectively).

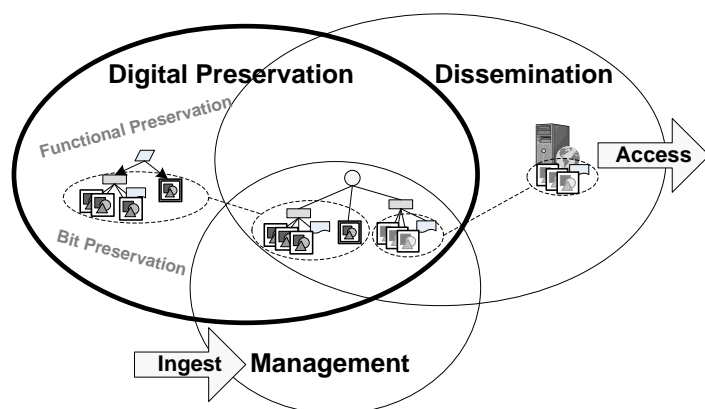


Figure 4 Digital preservation in an environment with dissemination

A *representation* consists of files of different formats and structures between the files or fractions of the files. An example is a book which can be represented by images of book pages along with digital information describing the order of the pages. In OAIS, *Representation Information* is defined as the information that maps a composed set of bit sequences into more meaningful concepts [16]. Here the full representation includes all elements that enable the meaningful presentation of the digital material. A representation in PREMIS is defined similarly as the set of stored digital files and structural metadata needed to provide a complete and reasonable rendition of the intellectual entity associated with the object [132].

The representation for dissemination and the representation for digital preservation may be the same, but there are also cases where they will differ. The reason can be that dissemination requires less volume requiring file formats or file formats with special functionality e.g. small volume GIFs or pyramid-TIFFs enabling zoom in images. In such cases the dissemination and representation platforms must be integrated. This integration could be by direct access to the bit preserved material. It could also be two completely separate platforms, where dissemination representations are only derived from the preservation representation when they are created on the dissemination platform or in cases where errors have caused loss of the material from the dissemination platform. A third option is that the dissemination platform is based on a cache area where the most frequently referred to material is placed, and in cases of reference to material that is not in the cache, this material will be derived from the preservation representation.

Identifiers for different representations, and parts of representation, must also be coordinated. This fact can contribute to requirements on how identifiers in the preserved material are represented, how it relates to identifiers used in dissemination, and how information is ensured in order to re-establish such relations in case of major losses in the dissemination platform.

The explicit requirements coming from the use of the digital material will differ according to the intended purpose of the material. An example is that libraries must in most cases both ensure preservation of the digital material, and they must ensure dissemination to the public in a fast and user friendly way. Another example which would result in other requirements is that archives have parts of the data that must be

preserved and kept confidential for 70 years, and subsequently be made accessible to the public. Again for other businesses like the pharmaceutical industry, there will be more requirements related to their obligation to keep confidentiality, but at the same time they must be able to deliver urgently required documentation, e.g. in cases where the product is suspected of having life threatening side effects. There are many other examples, but the final example here is that research communities need to preserve and share research results on a daily basis in an international context.

As illustrated in the examples the context has importance for choices made for the representation of digital material that is bit preserved, and it will therefore be the subject of further presentation in chapter 3 “A Representation Concept - for Analysis of Bit Preserved Data”.

1.5. The State of the Art for Relevant Areas of Digital Preservation

During the last two decades, many important and major steps towards sustainable digital preservation have been taken. However, technological evolution has not been stationary. This evolution has provided means in the form of tools, storage technologies and processing techniques, but it has also resulted in greater challenges in the form of increases in data volume, increased complexity in some digital materials, and increased demands for structured and fast availability. There is still a need for much work to achieve sustainable digital preservation as addressed by the task force in 1996.

One of the main challenges is that preservation is dependent on ongoing financing, which has been pointed at in several papers and reports, e.g. [9,143,150]. Parts of these challenges have been addressed by openness of standards and practices and contributions to a technical infrastructure.

This section will first describe the state of the art of bit preservation followed by the state of the art of relevant standards and practices, technical infrastructure, sharing knowledge and results as part of developing sustainable and affordable solutions. Finally, there will be a summing up of some of the relevant needed research.

1.5.1. Bit preservation

Bit preservation has taken different directions. One direction has focused on preservation of the digital media on which the bits are stored. Another direction has focused on backups. A third one has focused on the risks that must be taken into account in relation to bit preservation. The state of the art of these approaches will be described in the following.

Longevity of digital media

Longevity of digital media was seen as one of the important factors for basic digital information [188]. Different initiatives have taken the path of producing less fragile digital media. An example is found in the paper “Long Term Migration Free Storage of Digital Audio Data on Microfilm” [56], which argues for preserving bits as dots on a microfilm that can last for at least 100 years. The current state of other digital media can be hard to give precisely. The reason is that there are limited investigations of the life-time of media that are independent of suppliers [154]. An overview for optical media can be found in “Optical media longevity – the X-lab” [125], which estimates longevity to be about 10-30 years. Test procedures and test of optical media can be found in e.g. [120,159,179]. As discussed below, there are similar challenges in specification of the life-time of magnetic media. Furthermore, especially for magnetic disks in servers, the technological evolution in making e.g. more energy cost effective solutions means that there are rather short life-cycles for disks which in many cases are about 5 years [156].

Reliable storage considerations have also included use of Redundant Array of Inexpensive Disks (RAID¹²). RAID is used to make discovery and correction of some disk errors [21]. However, RAID is not sufficient to ensure bit preservation [192], it has to be accompanied with other means of bit preservation.

Often measures have been used to express the longevity of digital media, and as a measure for the reliability of a system, e.g. when using RAID. Examples of such measures are mean-time-to-failure (MTTF) and mean-time-between-failures (MTBF). There are, however, known challenges in expressing measures for bit safety in general. Calculations of e.g. MTTF can be made in different ways and take different parameters into account. The trustworthiness and usability of this measure has been discussed in recent years. The paper “Disk failures in the real world: What does an MTTF of 1,000,000 hours mean to you?” found that the MTTF given by manufactures was far higher than their results from experiments [156].

The measures are also sometimes used for more complex solutions with more replicas of data. Examples are measures which based on a simulation model with more replicas involved can be found in “Preserving the Bits of the Danish Internet” [21] and “Archival Repositories for Digital Libraries” [23]. The paper “Bit Preservation: A Solved Problem?” describes in detail the problems with MTTF and similar measures, and concludes that these measures cannot be a proper measure for bit safety on its own [143]. This thesis will refer to *levels of bit safety*, which denotes an unspecified way to differentiate between different bit safety requirements, where the highest level of bit safety is when no losses are accepted.

Replication and backup solutions

Bit preservation is sometimes understood as a question of replication in the form of backup. *Backing up* refers to creation of *backup* copies of data and storing them separately from the original data, so that these additional copies may be used to restore the data if it is damaged or lost [143]. Unless there are only requirements of a very low level of bit safety, backing up is however *not* proper bit preservation. The reason is that errors may occur in a backup copy of data, and these errors can only be discovered if the backup copy is checked against the original. Should an error be discovered in such checks, then the question is which copy is the correct one. Posing these questions already points away from traditional backing up and towards *active bit preservation* which is here used to denote bit preservation where copies of data are equally worthy replicas that are actively checked for integrity and existence on a regular basis.

Examples of interpretation of backing up as a way to do bit preservation can be found in several present systems. One example can be found in the “Policy and Procedures for Digital Archiving at The Henry A. Murray Research Archive” [174]. Another example is the Hoppla system which is aimed at small institutions with obligations to do digital preservation [166]. Although the paper does acknowledge that bit-stream preservation is not a solved problem, it states that Hoppla includes bit preservation. However according to the user guide this bit preservation is based on backup [58].

Another example of a backup based bit preservation system is the Digital Information Archiving System (DIAS), which has been used in the German Kopal system [79], and at National Library of the Netherlands (Koninklijke Bibliotheek) [123], [30]. The DIAS system is an example of a backup system where a number of backups are taken of the original data [30].

¹² “Redundant Array of Inexpensive Disks” according to initial scientific work on which it is based [126], but later renamed to “Redundant Array of Independent Disks” [19].

Clouding services to support storage of multiple copies of data in different places also provides examples of replication of data, and is presently a popularly discussed solution. Different cloud solutions are more or less prepared to meet bit preservation challenges. An example of a pure cloud solution, solely based on replication to ensure bit safety, can be found e.g. for IBM Cloud Computing [61]. Later in this introduction, there will be a more detailed description of DuraCloud, which is a cloud solution that includes more aspects of bit preservation than just replication [34].

Risk based approach

Although bit preservation has been mainly left to hardware manufacturers, there has been some discussion in the past years which goes further than taking hardware and backup perspectives. In 2005 the paper “Why Traditional Storage Systems Don’t Help Us Save Stuff Forever” [4] describes a number of threats that must be taken into account when to achieve preservation, and many of the threats concern bit preservation. Examples of such threats are storage failures, human operations errors, attacks, loss of context (metadata), and economy constraints. In 2006 it was followed up in the paper “A Fresh Look at the Reliability of Long-term Digital Storage” [5] which also focuses on the threats and points at the importance of independent replicas, as well as frequent detection and repair of errors. The same year the paper “Long-Term Threats to Secure Archives” [165] presents threats including confidentiality issues. It emphasises the need to focus on active integrity check, and ensuring indexing of materials. In 2005 the paper “The Requirements for Digital Preservation Systems, A Bottom-Up Approach” presents a detailed list of threats and strategies to avoid threats which again focus on replication, media migration, independence between replicas, and also points at the importance of audits and economy [144]. Although this is not a complete view of the literature and thus evidence of literature addressing bit preservation aspects, the lack of focus on the problem was explicitly addressed in 2008 in David Rosenthal’s paper: “Bit Preservation: A Solved Problem?” where he indicates limitations in current practices, as well as pointing out that bit preservation may not be as solved a problem as anticipated by most institutions and companies [143].

Examples of solutions with some risk based approach

The DSpace system is a repository solution [33,160,168], which has often been referred to as a bit preservation solution [80,135,152,164]. Bit preservation in DSpace is based on retrieved metadata (file format, MD5 checksum, creation date) kept in a separate metadata store. Furthermore, it is based on production procedures (e.g., high-quality servers and storage devices, good backup and disaster recovery plans). This allows for more replicas of data, but the metadata is not secured in DSpace, and there are no descriptions of the mechanism for integrity checking across the multiple storage locations.

In 2007, a possible extension of DSpace was presented in the paper “Automated Validation of Trusted Digital Repository Assessment Criteria”. This paper describes how a DSpace system can be combined with iRODS to setup different checks including the periods for integrity checks [110]. iRODS (the Integrated Rule-Oriented Data System) is software middleware that enables specification of rules for digital objects possibly kept in multiple storage locations. Rule chains can be specified including activation of micro services, and thus define workflows defined by the rules [52,62]. *Micro services* are small procedures that perform a certain task and which are made available for the iRODS server code [103]¹³. The iRODS can thus add value to DSpace by assisting in ensuring regular bit integrity checks. However, it can only act on existing micro services, thus as long as there are no micro services for checking across the multiple

¹³ Micro services also relates to the micro service concept described in “An Emergent Micro-Services Approach to Digital Curation Infrastructure” [164].

storage locations, the system cannot discover threats of e.g. internal attacks changing both file and checksum.

The Chronopolis system is also referred to as a solution to handle bit preservation. It aims to meet requirements with three geographical dispersed replicas, identifying deterioration through curatorial audit reporting, and has mechanisms for replacement. However the identification of deterioration does only seem to be within each site having a replica [106]. The Chronopolis system has recently switched to use iRODs in the system to apply e.g. integrity rules [107]. As for DSpace, applying iRODs does add value, but not necessarily enough to avoid all the described threats to bit safety.

An example of commercial systems is the Silent Cubes platform which includes independence of different sites for replicas with respect to technology and geographical placement [40]. Another example is the commercial Ex Libris Rosetta system, which is an extension to the Aleph system that originally was designed for library staff to create and edit bibliographic records. The Rosetta version includes preservation, but it does not tell how bit preservation is made, and therefore it is not convincing in its documentation [36]. There are also other aspects to consider, if bit preservation is based on commercial systems. Firstly, there is a question of whether the responsibility of bit preservation can be outsourced to a vendor, and whether this is acceptable. Secondly it has a disadvantage in having the community around commercial systems defined by the vendor.

Examples of solutions with more advanced risk based approach

There have been some initiatives that have taken a more risk oriented approach to bit preservation, that focus on how risks for loss of bits are prevented. These examples take approaches of ensuring bits by several replicas of data, bit integrity check of replicas, and independence between replicas.

One example is the LOCKSS system (Lots Of Copies Keep Stuff Safe) [139]. The LOCKSS system is a peer-to-peer system for the preservation of digital material in a decentralised digital preservation infrastructure [3]. A *peer-to-peer* system typically lacks a dedicated, centralised infrastructure, but depends on the voluntary participation of peers to contribute resources out of which the infrastructure is constructed [153]. The initial version of LOCKSS came in 2000 [140]. It consists of a number of independent LOCKSS caches that cooperate to detect and repair damage to their content by voting in opinion polls. The LOCKSS system is used widely among libraries and publishers. There are several examples of recent initiatives, which use LOCKSS as basis for Private LOCKSS Networks. *Private LOCKSS Networks* (PLN) are networks of specific LOCKSS caches defined by a small private community sharing a LOCKSS network [134,138,139]. The examples can be found in the MetaArchive initiative in the United States [158], the UK LOCKSS Pilot Programme in United Kingdom [151], and in Kopal which previously was based on DIAS [157]. There will be more about LOCKSS in chapter 2 “The IR-BR Model - for Analysis of Separated Bit Preservation” and chapter 4 “A Bit Preservation Evaluation Methodology - for Choice of Solution”.

Another example can be found in the archive part of the Danish web archive [20,116]. This is not a peer-to-peer system, but builds on the same bit preservation principles as LOCKSS; having independent replicas and regular active integrity checks [117]. However, this system is only designed for a specific solution with fixed number of replicas and fixed placement of the replicas, fixed security level when data is passed to and from the archive etc.

A newer example is the initiative around DuraCloud [34]. DuraCloud leverages existing cloud infrastructure to enable durability and access to digital content based on replication and integrity checking across multiple cloud providers. DuraCloud has with NDIIPP in 2009 launched a one-year pilot

program to test the use of cloud technologies to enable access to digital content [90]. Results from the pilot are only available from a PowerPoint presentation, which mainly focuses on technical issues. However, use of cloud for replicas should be done with care if the digital material in question has confidentiality issues. Furthermore, evaluation of how well a cloud solution meets requirements for bit preservation has some challenges, which will be discussed in more detail in chapter 4 “A Bit Preservation Evaluation Methodology - for Choice of Solution”.

A general view of a bit preservation solution

Common to all bit preservation solutions is that they include replication of the digital material. The difference between the solutions is the way that the replicas are placed on units consisting of a specific media in a technical and organisational environment. In the rest of this thesis such units are called *pillars*. It is also common that there are some sort of general system to ensure coordination between the replicas, e.g. to support replication at ingest and possibly to support integrity checks and operations.

Throughout this thesis, it is assumed that a bit repository can be viewed as illustrated in Figure 5, which expresses the assumption that bit preservation will include replication and coordination between replicas.

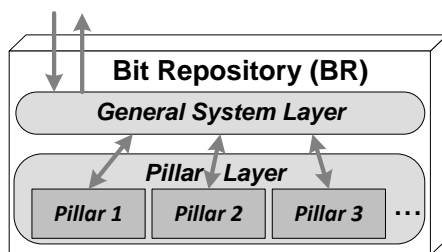


Figure 5 A general view of a bit repository

The arrows which point into and out of the BR represent ingest and access of bit-streams. The arrows between the *General System Layer* and the *Pillars* represent the connection between them. The pillars in Figure 5 can optionally have one or more replicas of specific data, and the general system layer includes all needed coordinating activities.

In the general view a tape based backup system for a server would have two pillars; one for the tapes and one for the server. The general system layer for the backup system would for instance contain coordination of performing the backing up as well as the needed data transmission in connection with the backing up. Another example is LOCKSS, where LOCKSS caches in a Private LOCKSS Network would be the pillars in a general view. In a peer-to-peer system like LOCKSS the general system layer will be thin, in the sense that a lot of the coordination and functionality is left to the peers. Still the communication protocol and network that enables coordination between replicas will be needed in this layer, which also will include coordination of components in the system in order to enable LOCKSS caches to communicate.

Requirements for bit preservation apart from bit safety

There are several examples in the literature that mention requirements for bit preservation other than bit safety requirements. For example access to bit preserved material via mass processing is mentioned in “Overview of the Netarkivet web archiving system” [22]. Another example of implicit combination of safety and availability is part of the preservation levels for the new Dutch e-depot [183]. Other examples are concerned with economy, for example discussed in the PhD thesis “Archival Repositories for Digital Libraries” [23], and focussed on in a model for preservation costs given in “Cost Model for Digital Curation: Cost of Digital Migration” [77]. Confidentiality is mainly treated separately e.g. in the ISO 27000

series [68]. Some of the confidentiality issues can e.g. be dealt with using encryption. An example on computationally efficient techniques for confidential storage and transmission can e.g. be found in “Confidential storage and transmission of medical image data” [121]. There will be more about various requirements other than bit safety in the description of the main results of this thesis.

1.5.2. Standards and practices

A lot of work has been carried out on standard and practices, which often have been developed and anchored at archives and libraries internationally. One example is the US Library of Congress which hosts e.g. metadata standards such as METS and PREMIS [89]. Another example is the National Archives of the United Kingdom which hosts the PRONOM technical registry [111,176]. A third example is The California Digital Library (CDL) which, for instance, has done much work on preservation formats [15]. And a fourth example is the Library and Archives Canada which has made File Format Guidelines for Preservation and Long-term Access [91].

The number of developed standards and practices are many. This is both because of the rapid evolution within digital preservation, but also because digital preservation covers many disciplines across different sciences such as computer science and curation, and across many areas within computing. This thesis will give some examples of such standards that are relevant for a holistic approach to bit preservation.

An example of an area with several contributing standards is file formats. Standardised file formats are important for digital preservation, since standardisation can support long-term preservation [91]. An example of a standardised format for preservation is the PDF/A format [65,67]. File formats of archived files are also of great importance for later retrieval of the archived material. An example of a standardised storage format is the WARC format which originally was aimed at web archived material [66], but also can be used for other types of materials [157].

Examples of major contributions with relevance for this thesis are the OAIS reference model, metadata standards and auditing frameworks. These will therefore be described in more detail in the following.

General reference model

A well established and commonly used reference model for repositories with preservation is the OAIS Reference Model, which also has become an international standard. The OAIS model was originally developed for the space science community, but the model has broad applicability. The OAIS model provides a frame of reference in which it can assist in balancing the need for digital preserving and the need to keep pace with changing IT [177]. An overview of functional entities in the OAIS reference model is illustrated in Figure 6 below [16].

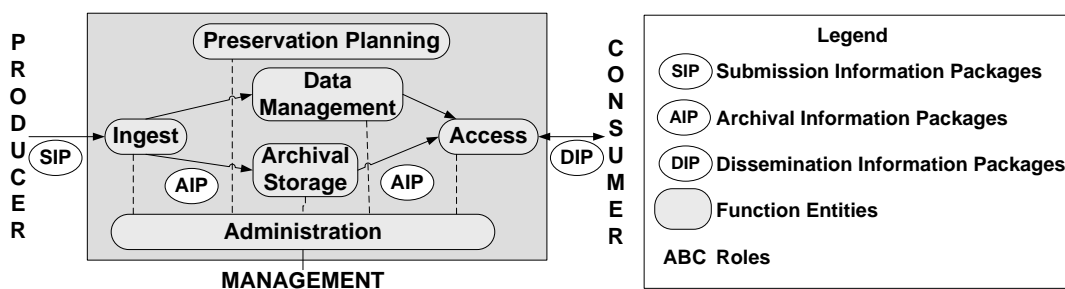


Figure 6 Functional entities in the OAIS reference model

The OAIS is very usable for reference and definition of terminology, and as the basis for analysis of existence of different functional entities in a system. Most digital preservation systems aim at being OAIS

compatible, a few examples are OAIS compliance of the KOPAL and LOCKSS system [42,79]. The OAIS reference model does not aim to describe actual implementations, and thus is at a level of generality that does not suffice for actual implementations [177].

The OAIS reference model is basis for the IR-BR model described in chapter 2 “The IR-BR Model - for Analysis of Separated Bit Preservation”.

Metadata

There is a wide range of metadata standards at least as de facto standards. One reason is that there are different types of metadata. Another reason is that technical metadata has individual schemas for different types of file formats, for instance, still images and texts. Furthermore there are also different metadata standards covering the same thing, for example the XFDU standard [18] is a more OAIS oriented alternative to METS.

Especially METS and PREMIS have had importance in for steps towards sustainable digital preservation. METS is a widely used metadata standard, which includes all types of metadata [24]. PREMIS is specifically designed to contain preservation metadata, therefore a short description of the background for these formats is given here:

- *METS – Metadata Encoding and Transmission Standard*

The METS schema is a standard for encoding descriptive, administrative, and structural metadata regarding digital objects. The standard is expressed using the XML schema language and it allows inclusion of other metadata standards based on XML (which most other metadata standards are). METS was drafted in 2001 based on work on another schema, which aimed to create a proposal for a digital library object standard by encoding the mentioned metadata types. The Library of Congress maintains the official METS website [24,102].

- *PREMIS - Preservation Metadata: Implementation Strategies*

An international working group named PREMIS was established in 2003 by OCLC and RLG. The charge of the PREMIS working group was, among other things, to define an implementable set of core preservation metadata elements, with broad applicability within the digital preservation community, and to explore opportunities for the cooperative creation and sharing of preservation metadata. The work resulted in the PREMIS Data Dictionary version 1.0 in 2005. PREMIS covers preservation metadata on objects (e.g. identifier), preservation events, agents associated to events, rights, and relationships for associations between e.g. objects agents and events. Today the Library of Congress maintains the official PREMIS website [131,132].

There are many different ways to integrate METS and PREMIS schemas, since there are overlaps between the contents of these schemas as well as other schemas. One example is that metadata like a file format can both be seen as technical metadata and as provenance metadata, and the file format can be specified both in the PREMIS schema, but also in e.g. the MIX schema. Another example is rights metadata in METS and rights metadata in PREMIS. More examples can be found in “A Checklist and a Case for Documenting PREMIS-METS Decisions in a METS Profile” [184].

METS and PREMIS will be used in the example given in chapter 5 “A Simple Example Using the Results” and in perspective in chapter 3 “A Representation Concept - for Analysis of Bit Preserved Data”.

Auditing

Auditing is part of work with trustworthy repositories. In this area, a number of checklists and methodologies have been proposed. The best known auditing methods are TRAC and DRAMBORA which are given a brief description below.

- TRAC - Trustworthy Repositories Audit & Certification

The work with TRAC was started in 2002 by RLG and OCLC which jointly published Trusted Digital Repositories: Attributes and Responsibilities (TDR). TRAC is based on this work together with a wide range of checklists coming from other projects and organisation. TRAC is, to a large extent, based in the OAIS reference model [122]. There have already been examples of repositories that have been TRAC certified [130].

- DRAMBORA - Digital Repository Audit Method Based on Risk Assessment

DRAMBORA Developed jointly by the Digital Curation Centre (DCC) and Digital Preservation Europe (DPE). DRAMBORA represents the main intellectual outcome of a period of pilot repository audits undertaken by the DCC throughout 2006 and 2007. It presents a methodology for self-assessment supported by a toolkit that builds on the TRAC, but adds risk assessment to allow repositories to set their own internally defined goals and assess the compliance to these goals. It encourages organisations to establish a comprehensive self-awareness of their objectives, activities and assets before identifying, assessing and managing the risks implicit within their organisation [27,32].

Another example is a separate checklist reflecting the German administrative requirements for repositories, which have been developed by The German Network of Expertise in long-term storage of digital resources (nestor), and at an earlier stage, nestor has contributed to the development of the TRAC checklist [115].

The TRAC and DRAMBORA play an important role in discussion of auditing repositories. As pointed out by “The Significance of Storage in the ‘Cost of Risk’ of Digital Preservation” TRAC and DRAMBORA correctly take a holistic approach to trust, but still need more details on storage [192]. However, they are still important in relation to auditing and evaluation of bit preservation and will therefore also be mentioned in chapter 4 “A Bit Preservation Evaluation Methodology - for Choice of Solution”

1.5.3. Infrastructure and knowledge sharing

Over the years there has been a focus on sustainable and affordable solutions [53,146]. This is partly achieved through establishment of technical infrastructure and communities which have been provided by numerous initiatives in form of projects, national and international initiatives.

Contributions to the technical infrastructure

A range of different solutions in digital preservation have been provided through a wide range of projects and initiatives over the years. The tendency of most of these systems is that they are particularly aimed at specific challenges of either preservation or use of digital materials.

A range of solutions focuses mainly on dissemination of digital materials. This is for instance the case for most systems based on Fedora (Flexible Extensible Digital Object Repository Architecture) which is an architecture for storing, managing, and accessing digital content [41]. One example of such a Fedora based system is the EsciDoc system, which is aimed at scientific documents [137]. EsciDoc covers a wide range of issues for scientific data, but it does not include much on preservation. A similar Fedora based example is the Hydra framework for multi-function, multi-purpose, repository-powered applications [60]. Hydra covers a wide range of repository issues, but is also weak on preservation. The Ex Libris is an

example of a commercial solution which is mainly focused on cataloguing and seeks to cover preservation in the Rosetta system, but, as noted, it is not convincing in its documentation [36].

Examples of systems that are more preservation oriented are tools coming from the mentioned Planets and KEEP projects. However, these tools aim at functional preservation, and they do not cover bit preservation. An example of a system that covers partly functional preservation and bit preservation is the iRODS for rule based management of the digital material [186]. iRODS provides an important element, but neither functional nor bit preservation can be fully covered by iRODS, and iRODS is not concerned with dissemination aspects. Examples of systems that are specifically focused on bit preservation are the LOCKSS and DuraCloud systems, which both have their strength and weaknesses as discussed in part previously.

The reason for the differences in focus of the different systems is most likely that there are many different requirements for use and preservation of digital material of different types and levels of complexity. As will be described in section 1.5.4 “Need for further research”, there is a need for narrowly targeted initiatives to be consolidated in larger systems. Specifically, as described above, there is a need for systems that serve both preservation requirements and requirements for use. As it is part of an holistic approach to bit preservation to include both preservation requirements and requirements for use, there will be more on this topic in chapters 3 “A Representation Concept - for Analysis of Bit Preserved Data” and 4 “A Bit Preservation Evaluation Methodology - for Choice of Solution”.

Digital preservation for all on a national level

In the past five years there has also been a growing recognition of the fact that not all organisations that have the obligation to preserve digital material, have the resources, knowledge and tools to operate and maintain their own solutions. Thus a shared solution will have the potential economic benefits of a large-scale operation, and it will enable the possible consolidation of bit repository expertise, which cannot be supported individually by small institutions. This is for instance the case for many smaller archives and museums. In order to protect the cultural heritage it is important to offer sustainable solutions which enable such institutions to take preservation responsibility instead of outsourcing it to suppliers who cannot take the responsibility [158]. This has also led to different initiatives worldwide.

Systems aimed at institutions with little or no previous experience with digital preservation have been developed. Examples of such systems are the Hoppla system from Austria which focuses on simplified form for migrations [166], and the ARCHIVEMATICA which aims to meet requirements of low cost and technical complexity of deploying a comprehensive curation solution for third world countries [182].

In the United States an initiative has been taken to reach out to all institutions with interest and obligation to do digital preservation by establishment of the member based alliance National Digital Stewardship Alliance (NDSA). The alliance gives members access to expert information about current practices, tools and services [112]. Furthermore, the MetaArchive was established in order to find collaborative approaches to digital preservation services for repositories with cultural-related materials, [158]. This is a community-owned, community-led initiative comprised of libraries, archives, and other digital memory organisations that works cooperatively with the Library of Congress through the NDIIPP Program [100]. The MetaArchive offers support for bit preservation via Private LOCKSS Networks [158]. Chronopolis has also started cooperation with the MetaArchive to investigate moving data between a Private LOCKSS Networks and the iRODS based Chronopolis [105].

In Denmark an initiative has been taken to establish a Danish website digitalbevaring.dk, which aims to create a forum for knowledge sharing about digital preservation in Danish terms and on Danish conditions. It is aimed at Danish archives, libraries, and museums, which face the challenges of digital preservation. Specifically focused on bit preservation, there is also ongoing development of a National Bit Repository that is to offer bit preservation solutions for all cultural institutions in Denmark. The work on strategies for this National Bit Repository has been the basis for some of the research presented in this thesis.

There are other examples of national initiative that cover similar aspects, for instance the above mentioned German network nestor, and the Netherlands Coalition for Digital Preservation (NCDD) which is a coalition that covers digital preservation within the public sector of the Netherlands aimed at a sustainable technical and organisational infrastructure [118].

Sharing knowledge and results

As mention earlier in this introduction, knowledge exchange is done by many means e.g. knowledge exchange programs, research communities, conferences and publications, and national initiatives. This also includes communities for specific areas such as the International Internet Preservation Consortium (IIPC) whose goals are to enable collection and support in internet archiving and preservation partly by fostering development and use of common tools [63].

An important part of knowledge exchange is also the communities that have emerged around different solutions. An example of a community that has emerged from a specific project is the Open Planets Foundation (OPF). The OPF is a member based foundation, which has been established to provide practical solutions and expertise in digital preservation, building on the research and development outputs of the Planets project [124].

There are many examples of projects with communities formed around shared results from the projects distributed as open source. For example Fedora and Fedora based systems like EsciDoc and Hydra. Other examples are iRODS and LOCKSS, where LOCKSS specifically focused on easy and inexpensive technology for preservation purposes. The different open source based communities are important steps towards sustainable solutions, both in respect to economic issues and shared maintainable solution.

1.5.4. Need for further research

In May 2011, a workshop was held on shaping new visions for EU-research in digital preservation [173]. As input to this workshop, one report provided an overview on the research on digital preservation of initiatives co-funded by the EU programs [167]. This workshop resulted in a report which points at the need for more research, where following elements were addressed [12]:

- Need for the expansion and greater integration of communities of interest.
- Deeper engagement between the engineering disciplines within computer science and the arts and humanities disciplines
- Need to focus on automation and simplification, as well as the development of services to address increasingly complex digital objects.
- Emphasis on the development of well-expressed business models to support investment in digital preservation.
- Moving from small targeted initiatives in the early years towards consolidation of this work later on in larger integrated projects.

- Developing new approaches in the education of computer scientists to ensure that issues of digital preservation formed a core part of university curricula.

These are all important and relevant elements, which point at expansion and better coverage of digital preservation in infrastructures, knowledge sharing and better anchoring in different scientific environments at a higher level.

More specifically concerning elements in a holistic approach to bit preservation, there are many unanswered questions in relation to resolution for a final bit preservation solution. An often asked question is how many replicas of data should be included in order to give high level of bit safety [23,192]. The thesis “Archival Repositories for Digital Libraries” goes a step further and asks how frequently these replicas should be checked for corruption, and whether the replicas should be on a few expensive disks or many cheap disks [23]. The paper “Long-Term Threats to Secure Archives” mentions the need to take different threats and confidentiality issues into account [165]. In 2009 the paper “The Significance of Storage in the ‘Cost of Risk’ of Digital Preservation” specifically asks for more research on how to evaluate the risks against costs [192]. There are no finite answers to these questions, but the evaluation methodology described in chapter 4 “A Bit Preservation Evaluation Methodology - for Choice of Solution” provides an evaluation method to deal with aspects of such questions.

1.6. Contributing Results

The research coming from my work has resulted in a series of papers which form the basis of the three main results presented in this thesis. These three main results are all important results contributing to finding the optimal bit preservation solution, when the solution must be part of a whole, i.e. in a holistic approach to bit preservation. The three main results are:

1. ***A model for definition of separated bit preservation from other aspects of digital preservation***

This model is needed, in order to delimit what a bit repository must cover. It is the basis for terminology, analysis of functions within the bit repository, as well as analysis of the interface to the bit repository. There has not previously been any model to support analysis of a separated bit repository, thus the result is valuable for analysts who need to analyse separated bit preservation.

2. ***A concept for representations of digital material***

This concept is needed in order to make a proper analysis of the digital material to be placed in a bit preservation solution. The concept supports analysis of: how the representation must be for digital material to be bit preserved, which requirements exist for the digital material to be bit preserved, and what requirements lead to requirements for the bit preservation of the material. The existing representation concepts are either too general or too specific to support such analyses, thus the concept presented here will be valuable for persons responsible for digital preservation, who need to make specific analysis on bit preservation including both materials and solution.

3. ***A methodology to evaluate the choice between different bit preservation solutions***

The methodology is needed in order to evaluate what is the right bit preservation solution for specific digital materials. The methodology gives a structured way to choose between specific bit preservation solutions given specific requirements. The lack of, and the need for, means to choose the right bit preservation solution for specific bit safety, confidentiality and costs requirements has previously been identified [23,165,192]. This evaluation methodology is therefore a big step towards filling this gap, and will be valuable in decision making of bit preservation strategies.

The holistic approach sets a framework for bit preservation, which can support various requirements coming from the “whole” to preservation and use of the material. Within this holistic approach the aim is

to find the optimal bit preservation for specific digital materials. The three main results will contribute to this holistic approach by means of:

1. **Definition of a bit repository**, by use of the *IR-BR model* for separated bit repository (BR) within an institution's repository (IR).

The model assists in definition by separating the bit preservation task of digital preservation activities, at least on a conceptual level. That means a conceptual or concrete interface to a bit repository, which defines the delimited system that must meet the various requirements for bit preservation, thus it defines the scope of the optimal bit preservation solution.

The persons that will be in a need of such a definition will typically be persons who are establishing a repository which must include bit preservation facilities. Examples are libraries that must preserve digital legal deposit materials, research centres that need to preserve measurement results, a pharmaceutical company that needs to archive documentation of tests and production, local archives switching to digitised materials.

The IR-BR model is documented in Paper D. The perspectivation of the model result is provided in chapter 2 "The IR-BR Model - for Analysis of Separated Bit Preservation".

2. **Specification of various bit preservation requirements coming from the "whole"**, by using the *representation concept* to support analysis of the final form of the digital material to be bit preserved, respecting the various requirements coming from the "whole" including functional preservation and use of the material.

The requirements can be for the representation to be bit preserved (e.g. file formats, metadata and structure), for the digital material to be bit preserved (e.g. confidentiality, availability and bit safety requirements), and for the bit preservation solution (e.g. security of confidentiality, meeting access requirements, and giving required bit safety).

The persons who will need to carry out such analysis will typically be persons responsible for collections of digital material, or persons responsible for digital preservation in general. This will typically be a person with an IT background who can perform the actual detailed analysis and design of representations, but someone at the managerial level will have to decide which standards must be used as the basis for the representations.

The concept is based on Paper A which describes analysis of representations, Paper C which describes functional preservation aspects of a representation, and Paper B which describes how representations for digitised material can be influenced by preservation strategies and the process of making the digitisations. The perspectivation of the concept result is provided in chapter 3 "A Representation Concept - for Analysis of Bit Preserved Data".

3. **Means to find the optimal preservation solution respecting requirements from the "whole"**, where the *evaluation methodology* can be used to choose between bit preservation solutions, when as many requirements from the "whole" as possible are taken into account.

The requirements for a bit preservation solution can be based on the requirements for bit preservation. The evaluation methodology then assists in choosing among different possible bit preservation solutions which can meet the requirements to different degrees.

The persons with interest in the final solution are persons at the managerial level to ensure that requirements regarding budgets and overall strategies are met in a final solution, but also persons responsible for collections or digital preservation will have interests in order to ensure the best solutions for the digital materials. As discussed later, a person responsible for offering bit

preservation solutions can also have an interest in how solutions can cover the needs for bit preservation solutions.

The methodology result is documented in Paper E, and the perspectivation of the result is provided in chapter 4 “A Bit Preservation Evaluation Methodology - for Choice of Solution”.

The next three chapters present the three main results: “The IR-BR Model - for Analysis of Separated Bit Preservation”, “A Representation Concept - for Analysis of Bit Preserved Data”, and “A Bit Preservation Evaluation Methodology - for Choice of Solution”. This will contain motivation and perspectivation of the results. In order to put the results in a full context for technical readers, the chapter “A Simple Example Using the Results” will present a small example. The example is based on explicit digital material that is to be bit preserved, and the example will refer to use of all the results presented in this thesis. Finally conclusions are provided in chapter “Conclusions” and suggestions for further work is provided in the chapter “Further work”.

This thesis will only refer to bit-streams preserved in a bit repository as files, i.e. it will not consider e.g. representations in databases. Using reference to files eases the description, but it is not a limitation as long as the elements are considered as delimited bit-streams.

As mentioned earlier, the holistic approach described in this thesis does *not* cover all phenomena that can affect the choice of a bit preservation solution. An example of aspects that are not included is bit preservation using green IT where ecological aspects of IT are taken into account [49,161]. Another example is that aspects of mass processing of bit preserved digital material will only be mentioned. However, it should be noted that it would be possible to extend the results presented in this thesis, at a later stage, with both green IT and mass processing aspects.

2. The IR-BR Model - for Analysis of Separated Bit Preservation

The IR-BR model is presented in Paper D: "Cross Institutional Cooperation on a Shared Bit Repository". The model provides a terminology and basis for analysis of separated bit preservation in an institution's repository. The IR-BR model includes the perspective of a bit repository possibly shared between several institutions, but the model can generally be used for systems where bit preservation is seen in separation.

In a holistic approach to bit preservation, bit preservation must be delimited and defined at least on a conceptual level. The IR-BR model contributes to this by giving the basis for defining terminology and for analysis of a repository with separated bit preservation.

One example of the use of the IR-BR model is part of the paper. This case study points at the need, and proposes a suggestion, for a flexible architecture that can meet differentiated requirements. Furthermore the case study points at a need for ways to find and express specific requirements for a bit repository. Another example is given here which will be on repositories using a Private LOCKSS Network based system as a bit repository within an institution.

2.1. Motivation

The research which led to the IR-BR model was motivated by the need for a model to delimit and analyse a separated bit preservation solution. Delimitation in terms of scope and analysis based on well-defined terminology is a prerequisite for finding an optimal bit preservation solution. As this was specifically needed in a project defining an overall Danish national strategy for bit preservation, the research resulting in the IR-BR model was partly initiated by this project.

The aim of the strategy project was to define an overall strategy for a Danish bit repository (DK-BR) for Danish cultural institutions. The work was based on an evaluation of literature, international preservation initiatives and systems, and interviews with various libraries internationally. My involvement in this project was partly based on my previous role as responsible for the archive module of the NetarchiveSuite software used for the Danish web archive [117], and I was one of the main architects for the strategic architecture of the DK-BR.

At an early stage, the project decided to base terminology on the OAIS reference model. The reason was that all institutions participating in the project were familiar with the OAIS terminology, and were aiming at an OAIS compliant solution, which was to include the DK-BR for bit preservation purposes.

It quickly became apparent that it was *not* clear how the OAIS terminology should be used in a context with a separated and shared bit repository. For the architects, it was clear that the bit repository was much more than just the OAIS functional entity *Archival Storage*, but for storage operators it was less clear. In all cases it was not clear what parts of OAIS that had to be included. This led to actual research to find a proper framework for the definition of terminology that was missing.

As to this, the literature on bit repository solutions is scarce. The only reference model that could be of help was the OAIS reference model. Most of the other literature found is on hardware solutions or bit preservation solutions. The LOCKSS system was one of the most advanced systems for bit preservation found. The evidence that a bit preservation solution can benefit from being viewed as a full OAIS repository is implicitly evident from the LOCKSS documentation, which includes a formal statement of conformance to the OAIS standard [42].

The fact that there were no existing reference models that could be used for a bit repository may be explained in the storage focused approach to bit preservation, where the risk based approach did not give any basis for a reference model either. One attempt to use OAIS on preservation storage is documented in “The need for preservation aware storage: a position paper” [37] and “Preservation DataStores: Architecture for Preservation Aware Storage” [38]. As stated in [37]:

“Today, the OAIS model assumes traditional storage as its underlying archival component and relies on other components of the data management system to provide higher-level functions such as packaging of the data or the computation of provenance and fixity.”

The paper is focused on the metadata for preservation actions as part of the storage device. Thus there is in this case no need for other than the OAIS reference model.

Since OAIS was the only relevant reference model that could contribute to the work of the bit repository strategy, and taken into account that the project partners aimed at bit repository solution that could be part of OAIS compliant systems, the draft of the OAIS based IR-BR model was a natural basis for this research.

2.2. The IR-BR Model

The IR-BR model separates bit preservation in a bit repository (BR) as part of an institution repository (IR). This is a model for analysis of separated, and possibly shared, bit preservation. It is based on the OAIS functional model and illustrated by a case where different institutions with different requirements for bit preservations are designing a common bit preservation solution. Some of the major strengths of the model are that it is the first of its kind to define terminology for separated bit preservation as part of a full repository, and that it is based on a well-known existing reference model. Figure 7 illustrates the IR-BR model contra the traditional OAIS reference model.

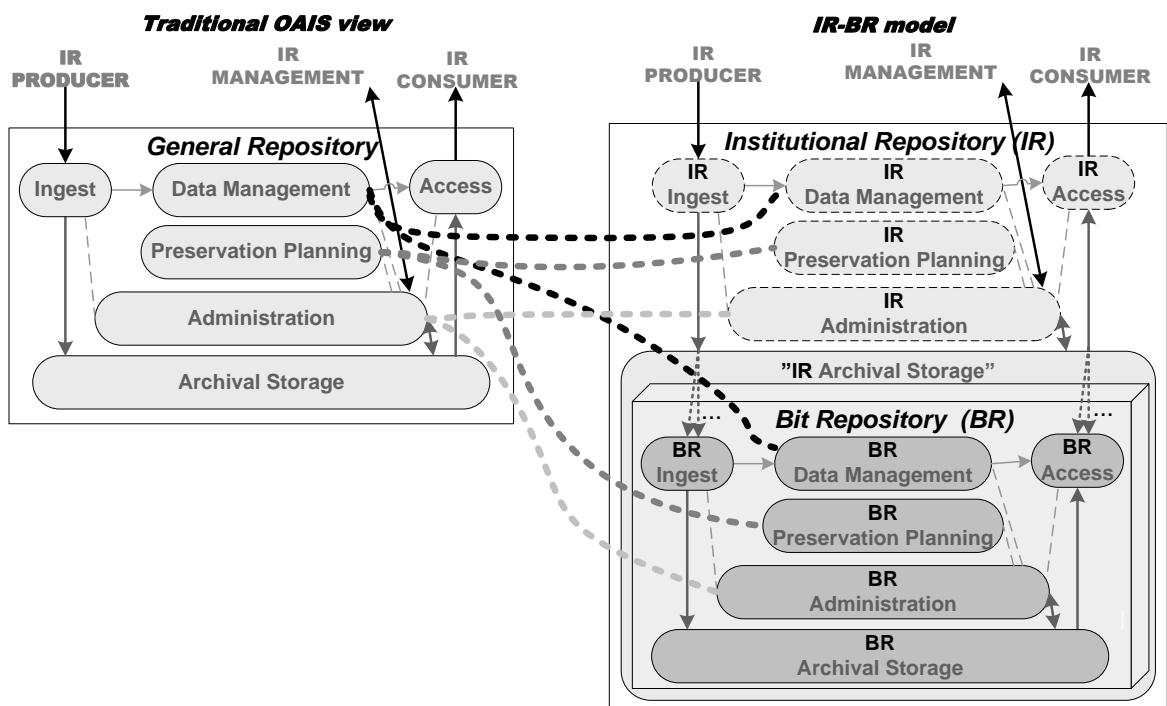


Figure 7 The IR-BR model contra the traditional OAIS reference model

The content of the *traditional OAIS view* to the left in Figure 7 is the same as the “Functional entities in the OAIS reference model” provided in Figure 6 in the introduction. The difference is that the functional

entities are moved around in Figure 7 to fit into an IR-BR model. The bold dashed lines indicate how the functions from functional entities in the traditional OAIS view are mapped into IR and BR functional entities in the IR-BR model. The bold dashed lines are given in different grey-tones for each entity in order to keep them illustratively separated.

The IR-BR model describes the bit repository as a repository within the institutions' repositories. The change compared to the traditional OAIS model is that the *IR Archival Storage* functional entity gets a new interpretation as a separated bit repository which includes parts of the general repository functional entities. The model enables description and analysis of the bit repository as an OAIS repository with all OAIS functional entities. This helps in analysis of what a bit repository must include when viewed as a repository in isolation. In this sense it serves to define what is involved in the bit preservation seen in isolation.

The main part of the analysis was to see whether the IR-BR model would violate OAIS on the overall level. Although the OAIS reference model has already proven useful for discussing issues related to long-term preservation, it is not strong in addressing bit preservation issues in a multi-organisational framework. It was therefore not obvious that the IR-BR model would not violate the OAIS reference model.

At the overall level, the result was that the IR-BR model does not violate the OAIS reference model. Only a minor redefinition was needed at the detailed level. The redefinition concerned a minor detail of the functions which communicate across the interfaces between the IR and the BR. This function had to be redefined in order to take into account that the data and information flow takes a different path. On the detailed level the redefinition of *IR-Ingest* function includes a redefinition of which function that delivers the BR receipt for accepted data and completed storage is returned to the IR. However, the receipt is given, which means that there is no violation on the overall level.

The fact that the IR-BR model makes sense within OAIS makes it a model which can fill the gap of the missing model for separated bit preservation, and its definition within an already well known reference model makes it even stronger. The analysis using the IR-BR model helps in defining bit preservation in the holistic approach to bit preservation. It also helped reveal locations of functionality and set up the requirements. Specifically it highlights a need for requirements that specify how audit trails within the bit repository are managed and possibly passed back, as well as definition of identification of the bit preserved data.

2.3. Use of the IR-BR model

The IR-BR model can be used for the repository of any institution where the bit preservation is defined as a separated part of the repository. Although 'institution' is primarily intended as public sector, i.e. educational, public service, cultural, there are no limits in the model that restrict its use by a private organisation, e.g. in the pharmaceutical industry.

Bit preservation will in most cases be in distributed systems in order to avoid risks of bit loss. In many cases the distributed system will cover placement of replicas in different organisations. The reasons for this can be that there can be risks for bit safety associated with having all replicas in one organisation, or that not all organisation have possibilities to achieve geographical distribution unless other organisations are involved. Thus analysis of organisational aspects will be relevant in many cases. This analysis can also be based IR-BR model, and will cover placement of responsibilities in the different organisations.

The Paper D on the IR-BR model included a case study which illustrates how the IR-BR model contributed to analysis and terminology. Additionally, the case study focused on design of a flexible solution that can offer differentiated services for different requirements. The most important points from the case study are briefly described below. Additionally, there is a discussion on how the IR-BR model can contribute to other repositories, e.g. based on Private LOCKSS Networks.

2.3.1. The case study from the paper

The case study described in Paper D was the project which examined an overall strategy for the architecture of a DK-BR. This project ran in parallel with the research done for the Paper D.

The IR-BR model contributed with important terminology and scoping for the bit repository strategy project, as well as analysis results focusing on identifiers, audit trails, and challenges of how to organise and make agreements with a DK-BR on the organisational level.

The analysis using the IR-BR model on a case study did not include an organisation around the BR. Therefore it did not point at placement of *Preservation Planning* concerning coordination of media migration for media containing different replicas of data. This is, however, needed in order to avoid too many simultaneous media migrations for different data replicas, which would add to the risk of loss of data. Another challenge in the architecture is to ensure *Data Management* data, e.g. how identifiers for delivered data are linked to data and how this information is secured. This is one of the points where OAIS is too general to help, so it remains a challenge for the final implementation.

An important point from the case study is that requirements differ according to the purpose of the collections of data, which e.g. includes the value of the data, requirements for confidentiality and access speed as well as costs. That means that the bit repository needs to meet differentiated requirements on bit safety, confidentiality, availability and costs. Furthermore, the different requirements could include e.g. political requirements to have at least one replica of data within the institution. Such political requirements can challenge the flexibility even further.

The mutual results from the IR-BR model and the analysis of requirements lead to the architecture illustrated in Figure 8. Here the *Clients* and the *Coordination Layer* belong to the *General System Layer* from “A general view of a bit repository” in Figure 5. The *Client applications* represent applications to perform operations like ingest, access and integrity check on pillars via the *Coordination Layer*. The pillars include illustrations of the platform they could be based on as examples.

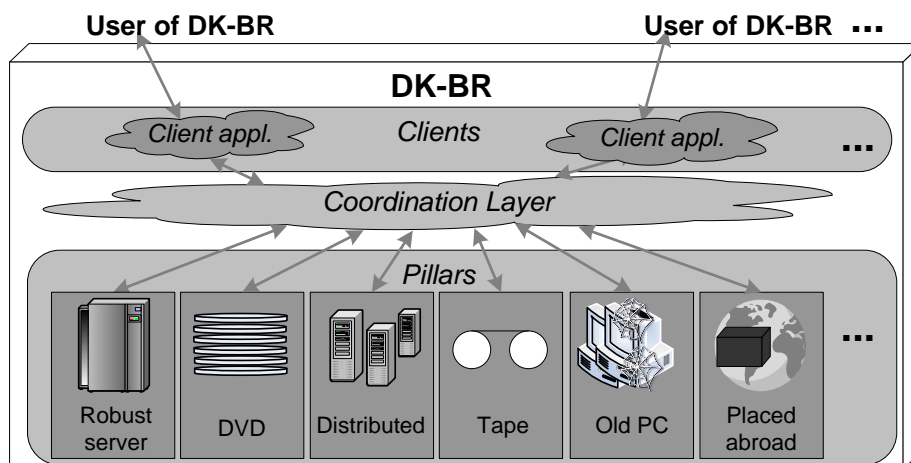


Figure 8 Architecture of the new Danish National Bit Repository (DK-BR)

The flexibility to meet different requirements on bit safety, confidentiality, availability and costs is created by the opportunity to have a flexible number of replicas where each of them is stored on a selected pillar. For instance data which needs fast access may need a replica on a distributed server solution, and data needing high bit safety may need a replica abroad, unless there is confidentiality requirements that cannot be met by having a replica under foreign legislation, etc.

The IR-BR model primarily focuses on the interface between the IR and the BR, i.e. separation of the functional entities where parts of these must be covered within the IR and other parts within the BR. It does not cover the internal split within the BR. However, if pillars are operated and maintained in different organisations, there must be consideration of placement of e.g. technology watch for the media in use as part of the *Preservation Planning* for media migration. If this part of the *Preservation Planning* is placed in individual organisations, then the organisational relation and information flow to the *BR Preservation Planning* must be analysed.

Analysis of organisational interrelations in the case study is not only relevant for pillars, but also the client applications. This can give similar challenges e.g. concerning security aspects at *BR ingest*. Actually there are many different permitted scenarios of use of the DK-BR, where a BR-user can be host of pillars and client applications. The BR-user can also choose to have internal pillars that are not part of the DK-BR, but are part of the BR-user's bit preservation solution. Some of the different scenarios for use of the DK-BR are illustrated in Figure 9.

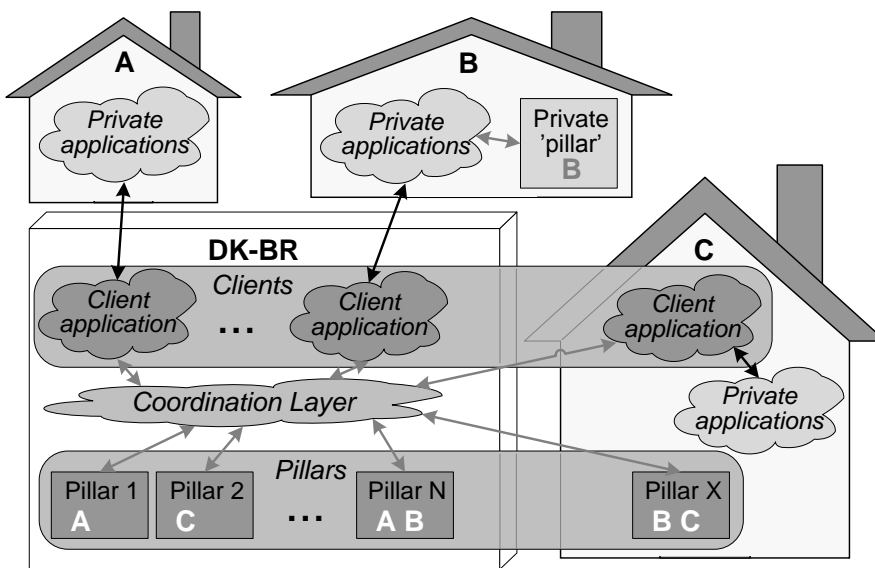


Figure 9 Scenarios for use of the Danish National Bit Repository (DK-BR)

In Figure 9, organisation A chooses to leave all bit preservation to the DK-BR, which corresponds to outsourcing of their bit preservation.

Organisation B, in Figure 9 chooses to have an internal pillar as part of their bit preservation solution. It will in this case be the responsibility of organisation B to carry out the bit integrity checks between replicas placed in the DK-BR and the replica on their private pillar. In terms of the IR-BR model, the BR for organisation B will be the full solution with the DK-BR as well as elements involved from B's private pillar application to ensure e.g. integrity between DK-BR replicas and private replica, and the involved internal organisation.

Finally organisation C, in Figure 9 has taken the choice of being host of both a pillar and operation of clients as part of the DK-BR. The BR for organisation C is, as for organisation B, will be a full solution consisting of both the DK-BR and internal elements.

One of the major challenges, with use of a BR, in an IR-BR perspective, is to find out what BR solution a BR-user needs to fulfil specific requirements, and how to express this in a service level agreement with the BR. A *service level agreement* (SLA) here means an agreement of level of service between the unit (s) responsible for the BR (the BR provider(s)) and a user preserving bits in the BR (a BR user). Such a SLA will need to express how the requirements are to be met by the BR; therefore the SLA will need to include details of pillars to some extent, e.g. physical location of replicas. It will, however, be preferable to have a SLA which is expressed on an abstract level without too detailed technical specification, since this will limit the ability to have independent pillar operation, such as upgrade of operating systems. However, the SLA must somehow express which changes in used pillars that have to be notified and possibly require renegotiation of the bit preservation solution, as for example in case of a media migration. Further work on these aspects is included in Paper E: "Evaluation of Bit preservation Strategies" described in chapter 4 "A Bit Preservation Evaluation Methodology - for Choice of Solution".

2.3.2. Using a Private LOCKSS Network as a bit repository

An additional example, of how the IR-BR model can be used, could be the KOPAL system's planned use of Private LOCKSS Networks for bit preservation. As mentioned KOPAL based on DIAS was argued to be OAIS compliant, and as mentioned the LOCKSS system has a formal statement of OAIS conformance. Thus it can be argued that a new KOPAL system based on LOCKSS is an obvious case of using the IR-BR model.

If the IR-BR model is used on a KOPAL system based on LOCKSS, it would require that the interface between the IR and the BR must be reconsidered compared to definition from the case study. The functions included in the BR's functional entities will depend on this definition of the interface. However, the interface must be definable with ingest and access points to comply with the IR-BR model. One possible definition of a *BR ingest* for a BR represented by LOCKSS could for instance be digital material are placed for ingest to LOCKSS (done via controlled harvest by the LOCKSS caches).

The Kopal system based on DIAS was argued to be OAIS compliant where the DIAS system was the *Archival Storage* of Kopal viewed as an OAIS system. Replacement of DIAS with LOCKSS will look different, since the LOCKSS is regarded as an OAIS system with *Ingest* and *Access* functional entities. Thus instead of storage and read of data from an *Archival Storage* in form of DIAS, the new system must store data by *BR Ingest* to LOCKSS, and read data by *BR Access* to LOCKSS. Thus use of the IR-BR model on this case can contribute arguments for OAIS compatibility.

The model can also contribute to terminology and analysis of separating the elements of OAIS functional entities placed in the KOPAL system and OAIS functional entities placed in the Private LOCKSS Network. Furthermore, LOCKSS caches in a network will be similar to pillars, in the case study, in the sense that they are placed in different organisations. Thus similar considerations of information flow to e.g. *BR Preservation Planning* will also be needed for a KOPAL-LOCKSS solution, i.e. which parts of the *BR Preservation Planning* is placed at individual LOCKSS partners and how it becomes part of *BR Preservation Planning* seen as a functional entity of the BR.

Private LOCKSS Networks initiated by MetaArchive will be a similar example of how the IR-BR model can be useful, since many of the functional preservation aspects are not dealt with in LOCKSS, which means these aspects of the preservation must be placed in systems using a Private LOCKSS Network.

2.4. Summary

The IR-BR model is the first model to provide terminology and basis for analysis of separated bit preservation in an institution's repository. The IR-BR model can generally be used for systems where bit preservation is seen in separation.

The usefulness of the IR-BR apart was illustrated in the case study, which also pointed at a need for ways to find and express specific requirements for a bit repository. Another example given was repositories using a OAIS compatible Private LOCKSS Network based system, which functions as a bit repository within an institution repository.

In a holistic approach to bit preservation, bit preservation must be delimited and defined at least on a conceptual level. The IR-BR model contributes to this by giving the basis for defining terminology and for analysis of a repository with separated bit preservation.

3. A Representation Concept - for Analysis of Bit Preserved Data

The representation concept is presented in Paper A: "Representation of Digital Material preserved in a Library Context". The representation concept has partly been developed on the basis of results on modelling the functional preservation aspect presented in Paper C: "Archive Design Based on Planets Inspired Logical Object Model". Furthermore development of the representation concept is based on the results from the results from Paper B: "Preservation of Digitised Books in a Library Context".

The representation concept presented in this thesis gives a nuanced perspective of relations between representations for different purposes and with possibly different significant information. The strength of the presented representation concept is that it supports analysis of digital material representations that must support different purposes in preservation and dissemination on a permanent basis.

In a holistic approach to bit preservation, the choice of representation of bit preserved material plays a central role. The choice must take into account which requirements there are to the future permanent access of the digital material, i.e. it must support later functional preservation actions, and it must be possible to fulfil requirements for use of the material, which includes means to identify the material on a permanent basis. Eventually, it must be possible for a bit preservation solution to respect all information security related requirements such as level of bit safety, availability, and confidentiality, as well as costs of bit preservation. Different choices of representation can affect the final requirements for bit preservation, since e.g. file formats can be more or less vulnerable to bit errors and require different storage volume. Thus requirements for bit preservation must be considered joint with the requirements for the bit preserved representation, which supports future preservation actions and use. The representation concept only gives the framework for analysis, whereas actual evaluation of the optimal choice of representations and bit preservation solution is covered in chapter 4 "A Bit Preservation Evaluation Methodology - for Choice of Solution".

3.1. Motivation

The research leading to the representation concept was primarily based on a specific research project on digital preservation in a library context, which was documented in Paper A and Paper B. The aim of the research was to get a better basis for definition preservation strategy for different types of digital material. The research question was whether it is plausible to reuse old digitised data in a normalised form. *Normalised form* here means a standardised form that minimises loss of information and ensures consistent use and maintenance of similar structures. The research relates to in what time frame the digital material must be preserved. Reusing digitisations will mean longer time frames, thus analysis of long term preservation and representation is relevant for such digital materials.

The resulting representation concept was also based on work at the Royal Library related to functional preservation, and investigation of enhancements of a digital object management system which had to include both preservation and dissemination. Work related to functional preservation was based on my participation in the Planets project, where specifically research in representations supporting functional preservation is documented in Paper C. The work, related to investigation of a common solution for preservation and dissemination, has contributed with many examples and highlights of the many challenges in combining preservation requirements with dissemination requirements.

A concept of representation can be found in different forms in the literature. As described in the introduction OAIS and PREMIS give very general definitions of representations. In other contexts,

representations are more operational and focused on specific challenges. One example can be found in “Using METS, PREMIS and MODS for Archiving eJournals” [25]¹⁴:

“A [representation] may be original or derivative, such as presentation copies or normalized preservation copies of the article. The [representation] object holds structural information about how its files relate and provenance information about the files' origin”.

This definition is closer to the representation concept presented in this thesis than the corresponding definition described in Paper C, where a presentation can be a migrations version, but it can also be a claimed new delivery of the material. Common for definition of representation in other literature is that it is either a very general definition, as it is the case in OAIS and PREMIS, or it is operational and used for a specific purpose. The representation concept in this thesis is operational, but takes a more holistic approach to representation for different purposes.

Other representation related terms used in the literature are terms of version, edition and generations. The meaning of these differs in different contexts, but in the general sense they are specialised kinds of representations [25]. In this thesis the representations for the same object are only called representations if they originate from the same material, digital as well as analogue. As described in the introduction there are also a wide range of challenges in connection with definition of how representations and their parts can be persistently identified. This is also a topic addressed in the representation concept.

The reason why there is a need for a new representation concept is the fact that the existing concepts are either too specific or too general as described above. This is probably related to the trend that repository solutions are either focused on use or on preservation as described in the introduction. Another reason is that the analysis of requirements cannot be done once and for all types of digital materials. There are many differences between collections regarding the legislation that needs to be followed, confidentiality issues, requirements for availability, but also regarding the complexities in the digital materials which can influence requirements for representations in order to support chosen preservation strategy and future availability. Thus the concept is needed to make structured and objective analysis that can result in a list of requirements for the specific digital materials.

3.2. The Representation Concept

The representation concept supports analysis of the required representations needed for different purposes and their mutual requirements to each other (Paper A). This includes structures and file formats for the digital material including metadata, as well as relations between them. Among the major strengths of the representation concept is that it provides a good basis for analysis of all requirements related to specific digital material, including use and preservation requirements.

Different representations are needed because their purpose dictates requirements for the form of the representation. An example with digital material that requires different representations can be found in publicly available library data, which must both be preserved and disseminated. As described in the introduction, requirements and technologies to support preservation and dissemination differ. Preservation needs a static representation in order to do the actual preservation. Dissemination has various requirements to be dynamic and quickly shift representation in order to serve the audience with

¹⁴ The paper uses the term manifestation for a representation (different to the IFLA definition of manifestations used in this thesis). In order not to confuse the concepts the terms manifestation in the quote is replaced by [representation].

use of the newest technology. Since they are representations of the same material they must be related and thus they will set different requirements to each other.

One example of requirements related to use is that legislation may require confidentiality for materials for a certain period, e.g. diaries from a deceased author. Another example is that digital material covered by deposit laws can result in requirements of high bit safety level. Likewise the use of the digital material can establish requirements of availability, where availability can include requirements for mass processing of bit preserved material. The functional preservation can e.g. have requirements to perform preservation actions on bit preserved material, which again can require a specific form the material. An example of a requirement related to actual functional preservation actions is to enable a one-to-many migration of many page image files migrated into one PDF file (Paper C). Another example is requirements of specific structure and presence of metadata e.g. identifiers and provenance metadata. The representation can also include considerations of choice and possibly the creation process of representations which can influence e.g. the ongoing costs of long term bit preservation (Paper B).

An example of different representations and some of their relations is illustrated in Figure 10. A more detailed small example of different implementations of representations can be found in chapter 5 “A Simple Example Using the Results”. On the left side the representations of an object are depicted along with some of the relations between the different representations, e.g. ‘*Latest Preservation Representation*’ is *migrated from* ‘*Earlier Preservation Representation*’. The right side of Figure 10 contains the legends and headlines for environment elements in Figure 4 presented in the introduction. The dashed lines between representations and environment elements illustrate to which environment element the representations primarily belong. The *Latest Dissemination Representation* belongs to *Management* because it includes dynamically *ingested* annotations that need to go through ingest procedures with e.g. quality checking before they can be made part of a new preservation representation. The *Latest Dissemination Representation* also belongs to *Dissemination* because the new annotation can be disseminated as soon as added.

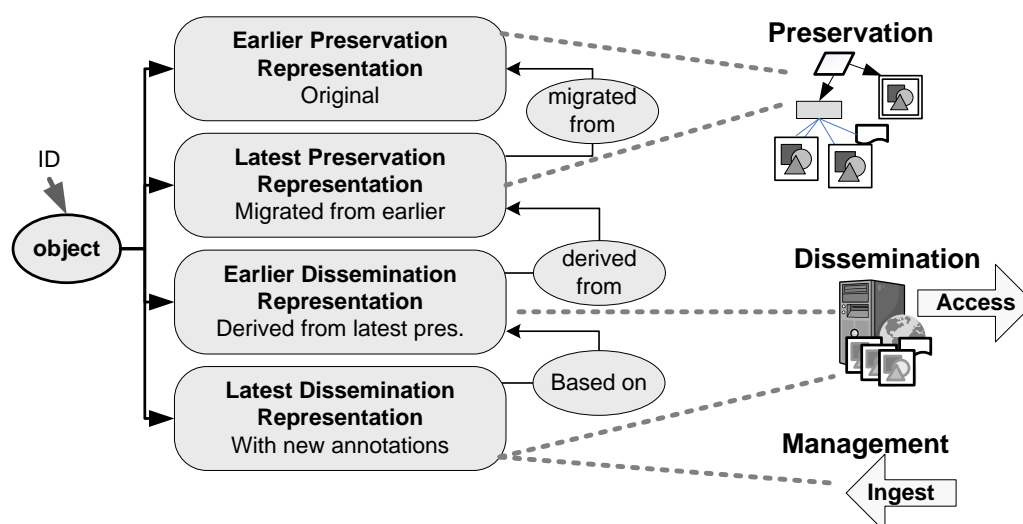


Figure 10 Example of representations of a book

Each representation has a purpose and the requirements for the representation must be defined from this purpose. Significant properties required for the representation must be defined by qualified persons with specific expertise, e.g. for preservation or dissemination. This includes required possible references

into a representation requirements dictated by dissemination tools, requirements for maximum preservation costs, requirements related to the value of the material and thus to bit safety level etc.

The representation description will to some extent need to be refined to possible implementation in order to analyse requirements like support of dissemination tools and cost of preservation. Possible implementation of representations will here include choice of file format, their quality as preservation and/or dissemination formats etc.

Some digital materials are more suitable for analysis using the representation concept than others, e.g. material from a broad crawl of the web may not need more representations, and it would probably require more work than the benefits gained if analysed via a representation concept.

A limitation in the concept is that the timing aspects that cannot be expressed directly in the representation concept. This is for instance the case if the preservation representation is migrated, then all relations and references will need to be updated at the same time. Another example is the challenge of annotations from the public [57], which may need quality assurance before being ingested and becoming part of the preserved digital material. Similar time issues are likely to appear for other digital material collections, as for example a global e-science network.

There will be different challenges which could result in possible extension in use of the representation concept. An example is deduplication on the file level. This would include considerations of modelling references to the physical files.

3.3. Use of the Concept for Bit Preservation Representations

The given representation concept is based on simple cases from the library sector. However, there are no specific restrictions in the representation concept which eliminate it from being used in other business areas with other requirements and thus a need for other representations for specific purposes. Common for all representation concepts in a digital preservation context is that there will be at least one representation which is a target for bit preservation.

In a holistic approach to bit preservation, it is in particular the choice of representation to be bit preserved that has interest. The representation for bit preservation must be defined in order to achieve the most optimal preservation and use. Thus this includes preparation for future preservation actions, and information security requirements related to current and future use. The choice must be optimal in the sense of meeting these requirements and at the same time respect requirements concerning costs of the full solution. The holistic approach does, however, only include aspects of digital material representation which are related to how the digital material is bit preserved.

The choice of representation for bit preservation will vary for different types of digital material, the purpose of the material, and the requirements coming from e.g. legal issues. It is therefore *not* possible to exemplify representation based on types of digital material; instead examples are structured according to types of requirements. Common for all are that a representation includes structures and file formats for the digital material including metadata, as well as relations to other representations aimed at different purposes. Below the examples describe how requirements can come from the “whole” and affect the choice of a bit preservation representation.

3.3.1. Confidentiality

Examples of requirements of confidentiality are mostly related to legal issues. Legal issues can for instance be copyrights or regulations that require some digital material must be confidential for a certain periods, e.g. diaries from a deceased author. Confidentiality requirements will result in requirements for how the bit preservation solution handles confidentiality of the bit preserved material.

Complexity of handling confidentiality requirements can also affect the representation of the bit preserved material. An example of this can be found in Paper A, where illustrations in public books were under copyright in a certain period. This will affect the way confidentiality is registered within the representation and it may also give extra requirements for how a bit preservation solution handles access to the material.

The specific requirements for how a bit repository handles confidentiality aspects will also depend on whether the BR-user decides to use encryption before placing the data in the bit repository. Use of encryption algorithms to encrypt the bit preservation material can, however, add extra risk of losing the material, and raises issues of how to handle bit preserved material in situations where the encryption algorithm is broken.

3.3.2. Availability

Examples of requirements of availability coming from the “the whole” are many. Availability requirements can set requirements for representation of bit preservations as well as requirements for a bit preservation solution. As described above, also access restrictions caused by confidentiality requirements must be handled.

A simple example is that dissemination requirements rely on the bit preserved representation. This could be the example of access to a web archive where an emulation strategy has been chosen. In such cases the preserved representation is the representation used for dissemination. Another example is dissemination based on a cache where access of material, which is *not* in the cache, will have to be derived on basis of the bit preservation representation within a specific time frame. A third example could be requirements for how fast lost data on a dissemination platform must be recreated on basis of the bit preserved representations. In all of these examples the bit preserved representation may be required to support derivation of dissemination representations including identification of representations. It is also common that the bit preservation solution will be required to deliver access at specified speed and load.

A more complex example from dissemination requirements is ability to access information that is calculated by means of mass processing, e.g. statistical information on web material. The bit preservation solution will in this case be required to support mass processing, and there will also here be specific requirements concerning access speed and load.

An example from dissemination that challenges relations between representations emerges from the increasing demands for the presentation of material, as for example the synchronous display of text in a book along with a video of an author recitation of the book described in Paper A. Depending on the role that the bit preserved material has for disseminated material, this can also result in challenges to access of parts of bit preserved material.

3.3.3. Bit safety

Bit safety is highly dependent on which bit preservation solution that is chosen for the material to be bit preserved. As pointed out for risk based bit preservation solutions, bit safety depends on the number of

replicas of the data, the independence between the replicas and the frequency of checking the integrity of the replicas. However, the choice of representation for bit preservation can also influence the bit safety, which is explained in the following.

Required level of bit safety can be influenced by e.g. deposit laws, where libraries have the obligation to preserve deposited materials. Levels of bit safety can also be influenced by preservation strategies and by the source of the digital material. As mentioned in the introduction digitally born material will for example in many cases require a high level of bit safety. Digitised material, however, as in the example given in Paper B, will only require a bit safety level sufficient to protect the economic investment in digitisation. Note also that in the case of protection of an investment, it is in particular the risk of losing all material that must be dealt with, since e.g. re-digitisation of a few pages is feasible, while re-digitisation of thousands of pages may be too expensive and time consuming to accomplish.

The choice of file format in a representation can affect requirements for bit safety, since some file formats like e.g. JPEG2000 are more sensitive to bit errors than for instance the TIFF file format [54]. As mentioned, compressed files are also sensitive to bit errors and choice of encryption can also affect the required level of bit safety, and the possibilities to ensure bit safety.

Finally, it may be valuable for future preservation actions, that the representation for bit preservation somehow includes information on the level of bit safety. One possibility could be to place the information in the metadata for the bit preserved representation.

3.3.4. Functional preservation

Functional preservation, in form of preservation action, must be possible on the bit preserved material. In order to perform such actions, there can be requirements for how the material is represented and how the digital material can be accessed.

A detailed example of elements that are crucial for the choice of representation as a basis for functional preservation is the file formats. Preservation file formats are usually required to be open and standardised, and there will often be requirements for the formats' ability to contain the significant properties that are chosen to be preserved for the specific material.

Functional preservation may also require that it is possible to do different kinds of mass processing on the bit preserved material. One example is retrieval of characterisation information via a tool. Another example is mass processing of a migration action within a bit repository.

Other important requirement for the representation of bit preserved material is structures in a bit preserved representation. An example of explicit analysis of requirements for a representation that must be used as basis for potential functional preservation like migration or emulation is given in Paper C "Archive Design Based on Planets Inspired Logical Object Model". This paper focuses on in particular the ability to do *many-to-many* migrations¹⁵, where an example of a many-to-one migration is many page image files being migrated into one PDF file. Furthermore it points to possible identification of the different representations.

Requirements for references into an object are in particular relevant for functional preservation, since referencing in many cases will be a significant property that must be preserved. This is relevant for the

¹⁵ More thorough description of many-to-many relations is given in the terminology chapter.

example references to citations, placement of annotations, and references needed for synchronous display of text and video. As illustrated in Paper A, there can be various complications in migrating references, which can also be the case for emulations, since referencing mechanisms will be dependent on each representation, and the available software to interpret the reference.

Metadata as part of the bit preserved representation is in particular important for functional preservation, since the metadata is the basis for functional preservation. It will in most cases also be the metadata that must include the description of structures in a digital object, and possibly how references into the digital object are defined. The level of preservation in terms of how functional preservation must be supported can also be part of the metadata.

3.3.5. Metadata

Metadata is a crucial part of the bit preserved representation. As mentioned above, the metadata is the basis for functional preservation, for identification of the representation and its part, possibly for registration of preservation level including level of bit safety, and possibly for access rights in connection with confidentiality. Furthermore, the metadata must include information about the preservation actions taken. In order to support as uniform treatment of bit preservation as possible, standardised forms of metadata are required. The introduction gave examples of metadata standards where especially the METS standard and the PREMIS standard have played major roles in standardisation of metadata to preserved digital material.

The METS standard explicitly has a part for structures of the object, e.g. pages in a book and the order of the pages. Structures can also be defined in MODS, which is one of the examples of how metadata can be specified in several places and result in risks of inconsistencies or inconsistent updates. Policies for specification of possibly redundant information in metadata for representations are therefore also important. This can partly be specified in profiles as discussed in “A Checklist and a Case for Documenting PREMIS-METS Decisions in a METS Profile” [184].

The level of preservation can be specified in PREMIS using the `preservationLevel` fields. There are no explicit guidelines for defining preservation levels. The paper “PREMIS With a Fresh Coat of Paint” discusses the definition of a preservation level in PREMIS, and explains that it must include both the level of bit safety and the level of functional preservation [85]. The challenge of a proper way to define bit preservation is therefore also a challenge to express specific preservation levels for the metadata. Furthermore, it is a challenge to express the preservation as combined levels of bit preservation and functional preservation. An example of very loose specification of preservation levels can be found in the DPC definition of digital preservation which is based on time frames of ‘indefinitely’, ‘defined period but not indefinitely’, and ‘defined period where needed technology is assumed to be available’ [31]. A more complex, but still loose, example is the definition of four preservation levels for the new e-depot at the Royal Library of the Netherlands [183]. The challenge of specifying preservation levels is, in all cases, to be sufficiently general to have preservation levels that can be used independently of changes in preservation solutions, and to be specific enough to ensure that it can be interpreted as intended.

A crucial basis for the metadata is that all needed identification information is present in the metadata. This involves identification of the representation itself, identification of the intellectual entity which the representation is for, and identification of the parts of the representation. All metadata standards include a specification of an identifier for the object they relate to. This covers identification from metadata to identifiable components of the representation. The PREMIS standard also includes the possibility of

identifying the intellectual entity in the field `linkingIntellectualEntityIdentifier` as part of the object information. A more explicit example is given in chapter 5 “A Simple Example Using the Results”.

3.3.6. Identification

There are different types of identifiers needed in a representation. There are the persistent identifiers on the logical level which must enable persistent reference to the material, no matter if the material changes as part of migration actions. Such identifiers will be referred to as the logical identifiers. There are also the persistent identifiers on the more concrete technical level, which point to specific representations or part of representations, e.g. files under bit preservation which are not allowed to change. Such identifiers will be referred to as the physical identifiers.

As mentioned in the introduction, a major challenge with the logical identifiers is that they work best if it can be assumed that an object has a linear version history. Another major challenge mentioned is the fact that the meaning of persistence cannot solely be technically defined, but will depend on social definitions including the purpose of the persistent identifiers.

One definition of the logical identifiers can be that that non-linear version history is not allowed, which will mean that new logical identifiers must be made for branches in the history. Otherwise the definition of a persistent identifier may be a basis for selection between representations. No matter what the definition a logical persistent identifier is, the definition will also define when a changed digital object is regarded as a new object with a new identifier, rather than a digital object which candidates to be addressable by the same persistent identifier. The representation concept cannot solve this, but can contribute to analysis of choices, which must be mirrored in the representations that are bit preserved.

One way of solving the challenge of logical persistent identifiers to branching version history was originally included in a full paper version of Paper C. This full paper version presented a model where the logical identifier could be qualified with a service which detailed which representation that was requested. Figure 11 illustrates this model, where the service provider is a repository, objects are digital material and object ids are logical identifiers to the digital material.

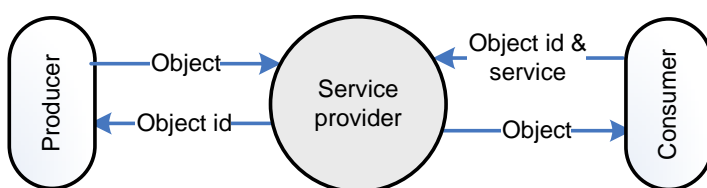


Figure 11 *Service based persistent reference*

A service could for instance be time based and/or based on significant properties. Implementation such service based persistent references would probably also influence the analysis of which metadata a representation must have in order to support such services.

It is noteworthy that no matter which decisions are taken for persistent identifiers, there is no problem in representing them in the metadata, since it is solely a matter of choosing the identifiers in order to support referencing by the logical identifier.

Concerning the physical identifiers, the challenge is how these identifiers are coupled with the physical files. A weak link in preservation can be to leave this coupling entirely to the bit repository, since there is

no direct evidence in the delivered bits on how they are identified. Most metadata standards are XML based which are not suitable for inclusion of files, thus e.g. METS would not be suitable for the coupling. Many file formats can contain an embedded identifier, e.g. TIFF, but extraction of these identifiers would then depend on the file format and this strategy cannot work for all formats. The coupling is often handled by use of packaging formats e.g. TAR, BagIt, or WARC [84]. Again the choice of format depends on what is viewed as acceptable risks. The TAR format assumes a file structure and does not support any identification mechanisms except from the file and folder names. The BagIt has similar challenges, but it allows specification of a single external identifier, i.e. it could be used to apply external identifiers for bags with one item. Using file names as identifiers will only be possible if the risks of doing so are considered acceptable. The risks related to file names are related to the challenges of different limitations and interpretation of file names on different operating systems. Furthermore, file names are not part of bit preservation, and reference will also depend on existence of a file system in the future. WARC is a standardised web archiving format which is designed for web material, but is usable for other material as well (e.g. it is also used in the LOCKSS based KOPAL project [157]). WARC can support coupling of external identifiers and files, and is the format used in chapter 5 “A Simple Example Using the Results”.

An important issue for identifiers in general, is the standard used for naming the identifiers. Especially for logical identifiers, there can be different considerations to take into account when deciding whether a logical identifier must be opaque. An *opaque identifier* is a non-humanly readable identifier. It may be preferable to have opaque identifiers, since this will remove one motivation for name changes [178], i.e. from a preservation perspective it is preferable to have opaque identifiers. On the other hand it is more user friendly to have humanly readable identifiers, i.e. in a dissemination perspective it can be preferable to have non-opaque identifiers. There is general consensus that identifiers at best must be universally unique in order to avoid clashes of identifiers. There exist various standards for identifiers, e.g. ARK [83], Handle [171], and URN [162]. In the example given in chapter 5 “A Simple Example Using the Results” URNs are used consisting of UUIDs [87].

3.3.7. Costs

There are many factors that influence the costs of bit preservation. As noted in the previous sections there can be different requirements for handling confidentiality and availability which can affect costs of the services required of a bit repository. Likewise for functional preservation requirements for actual performance of migration within a bit repository can have costs. Costs related to storage volume of the material that are bit preserved must be seen in combination with costs related to bit safety. The basis for volume dependent costs must be based on the number of replicas of the material where the costs can differ for each replica. Furthermore, the costs of ensuring bit integrity depend on the costs of calculating the required information for each replica and the cost of performing the bit integrity checks as well as the frequency of performing bit integrity checks. Thus the costs will normally be increased for extra data volume and for higher bit safety, since volume, higher frequency, and extra copies of data will have costs. Thus, storage volume influences costs, and all elements influencing bit safety will therefore also influence costs.

An example of storage volume required for a bit preservation representation can be found in two different decisions on preservation formats. At the Royal library of the Netherlands they chose JPEG2000 as a preservation format [46], while the Bavarian State Library chose the more storage consuming TIFF format as preservation format [46,81]. The main reason for the difference in decisions was that the Bavarian State Library was not concerned with storage costs. This evaluation included aspects of whether the file formats were suitable preservation formats. It did, however, *not* include evaluation of different bit

preservation solutions, which could be quite relevant, since JPEG2000 is more sensitive to bit errors than TIFF [54], and therefore a higher level of bit safety would normally be required for JPEG2000, and thus result in a more expensive bit preservation, which could affect the choice of a bit preservation solution based on costs.

Volume costs are not only dependent on representation and file formats. For digitised materials the storage volume can also depend on the chosen digitisation process, and thus there can be possible benefits from a prior analysis of the digitisation process. The digitisation process will e.g. include choice of scanners and their setup, and post processing of the scanning, and possibly choice of subcontractors to do parts of the scanning process. An example of this is presented in Paper B: “Preservation of Digitised Books in a Library Context”. This paper describes different representations for different choice of a digitisation process, where especially the results on storage space required for the different page images illustrate that choice of digitisation process can have impact on costs of bit preservation. The different sizes are illustrated in Figure 12.

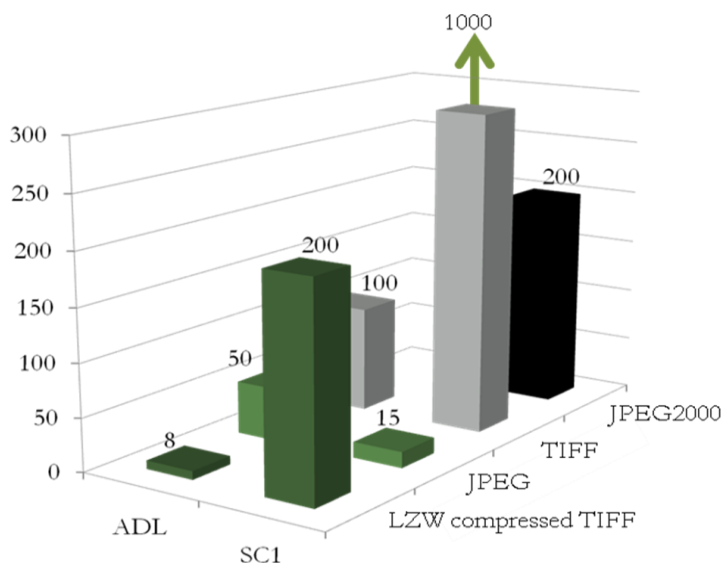


Figure 12 Storage space factor of digitised page images per file format

In Figure 12, ADL represents one digitisation process resulting in TIFFs (and with derived JPEGs), and SC1 represents another digitisation process resulting in TIFFs, JPEG2000 with lossless compression migrated from the original TIFFs, derived JPEGs and LZW compressed TIFFs. In spite of uncertainties related to these numbers the difference is evident.

The uncertainties related to the numbers in Figure 12 are partly related to differences for different pages where there may be differences e.g. due to the fact that some ADL images were done in black & white or done in grey tone. Furthermore, the sizes of SC1 TIFFs are based on TIFFs migrated back from the JPEG2000 which was needed because of an error in delivery from the SC1¹⁶. However, it is known that there is a general difference between JPEG2000 and TIFFs with at least a factor 2 [46], i.e. the pattern shown in Figure 12 would be the same for the original TIFFs, although the difference might not be as great.

¹⁶ Further description of SC1 is provided in the terminology chapter under description of ‘SC1’.

Some of the obvious reasons for the size differences are that the scanning results from SC1 were in colours while the ADL were in black & white or grey tone. Furthermore the images from SC1 include the full image of the page while the ADL had the margins cut off.

If influences of a digitisation process must be investigated, this could be done in form of a small pilot project, as it was done for the example from Paper B. Such a pilot could then be used to investigate which digitisation process to choose, based on the results of the pilot, analysis using the representation concept, and evaluation of bit preservation strategies given in chapter 4 “A Bit Preservation Evaluation Methodology - for Choice of Solution”.

3.4. Summary

The representation concept supports analysis of digital material representations that must support different purposes in preservation and dissemination on a permanent basis. This is the first concept to take this broader perspective, while at the same time it can include more concrete implementations.

The analysis of digital material representations is the basis for choosing representations and to find the requirements related to the representations. The choice includes the best choice of the representation to be digitally preserved. It can be used for any material in any context, although it is best suited for digital material where the representations can be described at a detailed level. For instance, the representation of web material can only be described at a general level.

In a holistic approach to bit preservation, the choice of representation of bit preserved material plays a central role in the search for the optimal bit preservation solution. The choice must take into account which representation is the optimal representation when taking into account the requirements for the present and future permanent access to the digital material.

4. A Bit Preservation Evaluation Methodology - for Choice of Solution

The bit preservation evaluation methodology is presented in Paper E: “Evaluation of Bit preservation Strategies”. The methodology provides a way to evaluate the best choice between different bit preservation strategies in the form of bit preservation solutions. It is motivated by the need for specification of requirements for specific digital material and evaluation of how the requirements are met. Furthermore, it will be needed in continuous evaluation of fulfilment of the requirements, and can also be used in choices of digitisation processes.

In a holistic approach to bit preservation, the choice of a bit preservation solution must take as many aspects from the “whole” into account as possible. The success of bit preservation depends on planning and carrying out the necessary preservation activities in time to fulfil preservation requirements. The bit preservation planning must be supported by tools which enable choice of optimal decisions for preservation [6]. The methodology presented here can be used as a tool to assist in taking such a choice. An optimal solution must fulfil various requirements for the bit preserved material, and requirements for bit preservation. The representation concept can contribute to express many of the requirements for the bit preserved material as requirements for a bit preservation solution. The methodology presented here is based on such requirements for bit preservation solutions and provides a way to choose between potential bit preservation solutions.

The presented methodology is based on the general view of a bit repository as a conceptually separated repository to take care of bit preservation. An evaluation is based on higher level requirements for such a bit repository. In the presented version, the requirements cover bit safety, availability, confidentiality and costs, where bit safety is formulated as requirements for how well the risks of losing bits are prevented. Furthermore, the presentation includes a description of tools that can support use of the methodology. The example of tools is most elaborated on bit safety aspects, but the example does also include some confidentiality aspects.

The methodology can be used for any bit preservation system which can be described in the general view of a bit repository. It can be used by providers of a bit repository as well as users choosing a bit preservation strategy for their digital material. Strategies do here refer to a specific solution in a solution space¹⁷ to a specific problem. That means a choice of a bit preservation strategy is a choice of a bit preservation solution. This chapter will include a description of relevant uses of the evaluation strategy, as well as discussion how a SLA can be defined between the provider and user representatives of a bit repository.

4.1. Motivation

As mentioned, in the section on the state of the art, there have been various occasion over the years, where the need for means of evaluation of bit safety has been pointed out. This has been related to how many copies are needed, but also evaluation involving confidentiality aspects and economical aspects. The evaluation methodology is motivated by some of the same aspects as the ones mentioned for these cases. It is a major step to choose between suggested solutions by evaluating which solution that best meets the various requirements for bit preservation. It overcomes the challenge of lacking well suited measures for bit safety, by expressing bit safety requirements more implicitly.

¹⁷ Theories on problem and solution spaces can be found in “System Analysis, Design and Development Concepts, Principles and Practices” chapter 14 [187].

The evaluation methodology was directly motivated by the work with IR-BR model, and also more indirectly by the representation concept. Especially the case study from the work with the IR-BR model left a question on how a BR-user can choose between differentiated solutions offered by a BR, and it left questions of what a BR must offer in order to cover a wide range of bit preservation services. The work with the representation concept left questions on how to choose between representations, when the different representations required different bit preservation solutions. The optimal solutions respecting other requirements like costs would therefore be to evaluate the representations together with required bit preservation solutions. Thus for both main results, a way to evaluate bit preservation solutions was needed.

There do exist different auditing methods like TRAC and DRAMBORA, which can be used for auditing a bit repository and evaluate how an existing solution meets quality requirements. However, these methods are not fit to make an explicit evaluation between how well more general requirements are met by different bit preservation solutions.

A tool like iRODS can assist in setting up rules that must be respected in bit preservation, but it is not a tool for actual implementation of bit preservation, and it is not useful for a total evaluation of the various requirements for bit preservation either. The Plato tool has proven useful for evaluation of functional preservation strategies [6], but it had not been used for bit preservation solutions previously. A starting point was therefore to see if Plato could support evaluation of bit preservation strategies. At an early stage of investigating this approach, it became apparent that there was a need for additional tooling, which led to my design and implementation of the Bit Repository – Requirement Measuring System (BR-ReMS) prototype. There will be a more detailed description of the BR-ReMS, including the reasons for design and development of the prototype after the description of the methodology.

4.2. The Methodology

The methodology enables evaluation of requirements for bit preservation that are independent of the chosen bit repository solution. The major strength of this methodology is that provides a way to express and evaluate how different bit preservation solutions fulfil requirements for bit safety along with other relevant requirements on a high level. Since there previously has been no methodology to support such choices, this evaluation methodology is an important step in the direction of choosing an optimal bit preservation strategy for specific digital material.

The methodology consists of specification and evaluation of high level user requirements in the Plato utility analysis tool. This evaluation is based on detailed calculation in the BR-ReMS of how well the requirements are met by different strategies defined in SLAs.

In the suggested specification of the user requirements for bit preservation, the requirements are expressed as scores relating to the probability of preventing a risk of losing bits. The choice of probability is made in order to have a uniform measure which can be expressed in different granularities. As stated in “The LOCKSS Peer-to-Peer Digital Preservation System”, describing evaluation of design to resist failures and attacks must necessarily be approximate and probabilistic [96].

The methodology is illustrated in Figure 13. Here the *Material Requirements* are specified in *Plato*. The *Plato* tool supports evaluation by use of *Utility Analysis* of how well potential bit preservation strategies meet bit preservation requirements. However, in order to evaluate the strategies, calculations must be done of how well each of the requirements is met by potential bit preservation strategies. These

measures are calculated by the *BR-ReMS*, which calculates the probabilities for fulfilment of the requirements for different bit preservation solutions which represents the bit preservation strategies. The *BR-ReMS* calculations are based on the various *Values of Characteristics* of BR solutions represented by a *BR implementation* and *Specific use of the BR* (SLAs). Thus, the requirements are specified in the *BR-ReMS*, as well as the different characteristics for each of the potential bit preservation solutions that the evaluation is to rate. The final Plato result of evaluation of how well the potential bit preservation meets the requirements is documented an *Evaluation Report*.

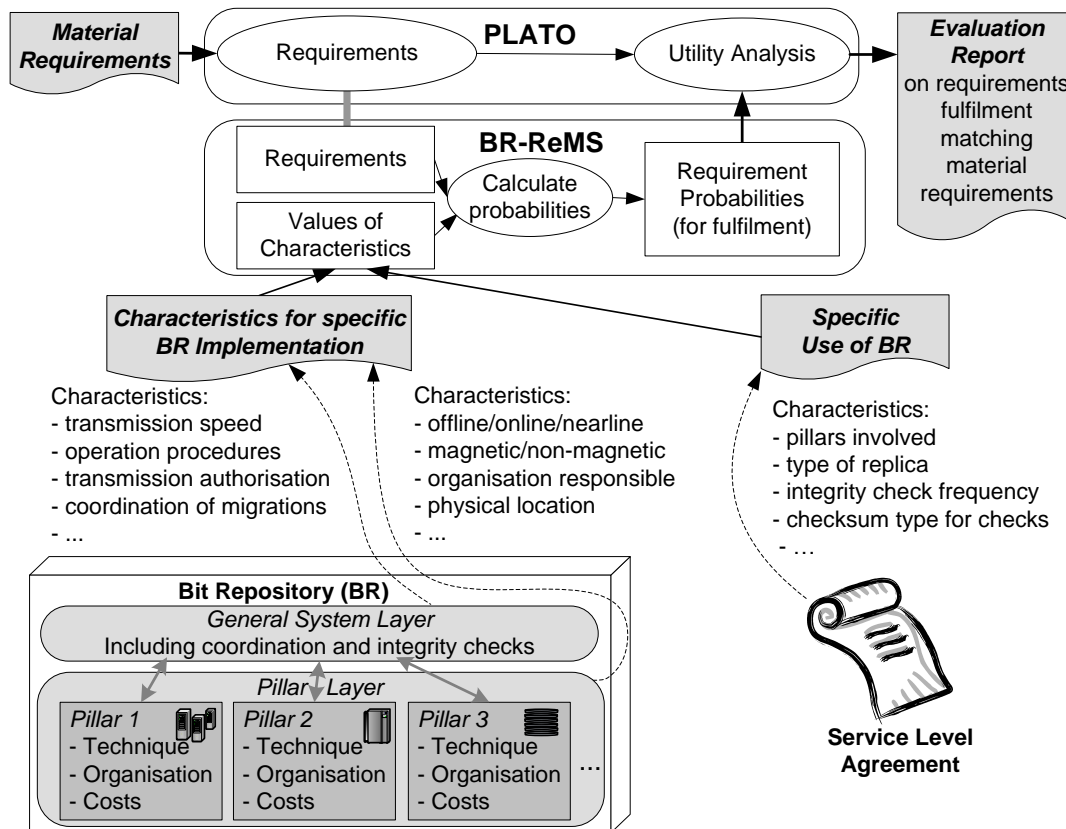


Figure 13 The bit preservation strategy evaluation methodology

There will be a detailed description of the different characteristics and calculations in the BR-ReMS in the next section. A small example of using the methodology, along with examples of characteristics and calculation of probability can be found in chapter 5 “A Simple Example Using the Results”. More extensive examples on materials and actual results from use of the methodology can be found in Paper D. Additional detailed use and output of Plato is given in Appendix I “Detailed Calculations using Plato”.

The BR-ReMS is an important part of the evaluation methodology. One of the main benefits of introducing the BR-ReMS as part of the methodology is that user requirements can be expressed in a form that are more easily understood. The requirements become simpler, since they become independent of the possible bit preservation solutions with details on technical, organisational and economic aspects along with the specification of how they affect the different requirements.

Another benefit from introduction of the BR-ReMS is that the methodology covers all kinds of different groups of requirements, even though the groups have conflicting elements. For example, bit safety can be increased by adding more replicas. On the other hand extra replicas also add costs, and extra replicas add risk of leaking confidential data. High and fast access to data also comes at a cost. Requirements of access

in form of mass processing can conflict with confidentiality issues on data placed in the same pillar. The BR-ReMS solves these interrelated conflicts by basing the calculations on characteristics. The characteristics are independent of the calculations that they are included in, and can therefore contribute to several calculations on conflicting requirements.

Introducing the BR-ReMS in the evaluation methodology does not solve the problem of defining bit safety, but it provides a way to break down the requirement into requirements expressing different parts of the challenges with bit safety, e.g. bit integrity check frequency and independence between pillars. It is also important to note that the calculations are not trivial or objectively definable for these sub-requirements either. For instance, there will be different opinions on whether bit integrity check on DVDs every second year is adequate. This is why the BR-ReMS also provides a possibility to define these calculations on characteristics in a way that can be presented to the user of the methodology, and thus the user can decide whether they agree with the way the fulfilment probability calculation is defined. This should include documentation for values of characteristics that are based on subjective measures. It is noteworthy that the methodology does not and cannot strive to give perfect results. The way to meet this fact is to enable users to review documentation in order to evaluate whether they agree with the way the BR-ReMS calculates results.

Although both Plato and the BR-ReMS have proven to be useful for the methodology, it is at the overall level independent of the specific tools. Plato could be exchanged with another utility analysis tool for evaluation of requirements. Another BR-ReMS could be used as long as it offers means of specifying the different types of characteristics and their values, as well as means to document the calculations of how well requirements are met, in order to enable possible review of the functions. In any case, the BR-ReMS presented here is only a prototype. A real implementation of a BR-ReMS must be extended with further specification of requirements and functions for calculation, as well as provide increased granularity of characteristics and calculation results. Furthermore, the methodology could be supported better by more advanced use of Plato, e.g. using prioritised requirements in terms of weights.

Detailed use of the methodology still needs some work. The requirements specification for the Plato and the BR-ReMS tools could benefit from further refinement. For example, the user requirements could be refined more. One example is that the requirements could state more clearly whether it is acceptable to lose a few bits but not acceptable to have mass loss of bits, which might be the case with digitised books that can be re-digitised. Another example is that the requirement to prevent loss in case of war or terror attacks could be split into one requirement concerned with war and one concerned with terror, since they represents two very different situations. The specified requirements could also be supplemented by a better breakdown based on the ISO 27000 series [68], and based on aspects from auditing methods like TRAC [122]. Furthermore, the requirements must be specified to include economy and availability. The economy part could for instance be based on the results from the preservation costs model¹⁸ described in the paper “Cost Model for Digital Curation: Cost of Digital Migration” [77]. The availability aspects ought to include requirements for access speed, and the ability to carry out mass processing close to the data.

Further extension of the methodology could involve new groups of requirements, besides bit safety, confidentiality, availability and costs, such as green IT, which may be part of future requirements for bit preservation [49,161].

¹⁸ The paper “Archival Repositories for Digital Libraries” [23] does also include economy considerations, but it is not as advanced as in “Cost Model for Digital Curation: Cost of Digital Migration” [77].

4.3. The BR-ReMS Prototype

The inspiration for my design of the Bit Repository – Requirement Measuring System (BR-ReMS) came from the prior top-down analysis of evaluating bit preservation strategies solely based on Plato. The Plato tool is based on analysis of requirements expressed in a requirements tree¹⁹, which consists of a breakdown of requirements into leaves with measurable requirements. Specification of a requirements tree for bit preservation turned out to be an almost impossible task. The difficulties came when various elements from details in a solution had to fit into a requirements tree for Plato. For example the number of replicas has influence on the bit safety, but so do facts about the exact pillars that the replicas are placed on, since independence between pillars influences bit safety. That means all relevant information on involved pillars somehow had to be presented as well. At the same time the choice of both number and specific pillars was actually part of the individual solutions among which choices had to be made. Furthermore, the number of replicas does also have impact on confidential aspects, which also would rely on specific pillar information. All these interrelated dependencies and details meant that we quickly ended up with lots of details that had to be taken into account in different parts of the decision tree. Such dependencies and interrelations cannot easily and comprehensibly be specified in a decision tree structure, since it has more the form of a network of relations rather than a tree of relations. On this basis we gave up the attempt of using Plato directly.

The first part of the Plato analysis did, however, lead to a requirements tree for a usable breakdown of information security requirements based on ISO27000. The breakdown of bit integrity requirements was expressed as how well a solution prevents different risks for bit loss, since this so far has been the best way to approach bit safety issues. The challenge of expressing numbers of replicas added to the confusion of specifying the requirements, and this resulted in the final placement as part of the expressed possible solutions among which to choose.

The described analysis of using Plato inspired me to split the task into two parts. One part included specification of requirements on a higher level which were expressed in Plato. This level of requirements was based on the initial breakdown using ISO27000 and risk based approach to bit integrity, since it seemed to have an appropriate and user friendly level, where details on technique, organisation and economy were left out. The other part was to calculate how the individual requirements were met by different bit preservation solutions. This was the part which led to the design and development of the BR-ReMS. The challenge was in this way met by dividing the problem of a full evaluation into smaller problems of how the individual requirements were met. Solving the smaller problems could then be basis for a full evaluation by use of the Plato tool.

The BR-ReMS includes many details of bit preservation solutions as well as a description of functions to calculate how the individual requirements are met. In this way, the split can also serve to hide the many details carrying out bit preservation from users of an evaluation system, although details of course must be available for users interested in the details.

The design of the BR-ReMS was based on a bottom-up approach. This was motivated by the prior analysis which had revealed that a lot of the elements affected how requirements were met. Thus the basis of elements that could be potential parameters for functions had to be the starting point. The analysis resulted in identification and analysis of the characteristics which could describe the influencing elements.

¹⁹ An example of a requirements tree can be found in section 5 “A Simple Example Using the Results”.

In my analysis of characteristics, I found that there were basically two main types of characteristics. One main type concerned characteristics of a bit repository offering different solutions, so called BR characteristics. The other main type was characteristics that were related to a specific use of the bit repository, i.e. specific for a specific bit preservation solution based on the bit repository. Since such a use of a bit repository would be based on an SLA between BR provider and BR user, these characteristics are called SLA characteristics. Thus each SLA only covers the pillars which are to have replicas of material covered by the SLA. Within these two main types of characteristics, there were two subtypes related to the general view of a bit repository illustrated in Figure 5; characteristics concerned with the system layer and characteristics concerned with pillar layer. That left four different types of characteristics:

- BR pillar characteristics
which are stable characteristics of a *Pillar* in the BR, e.g. media type
- BR system characteristics
which are stable characteristics of the *General System Layer* in the BR, e.g. data transmission protocol used for transmission of data from pillars
- SLA pillar characteristics
which are characteristics of a *Pillar* for a specific SLA, e.g. whether the replica placed on the pillar is a *full replica* or a derived replica in form of a checksum replica. A *checksum replica* is here defined to be a checksum of the replica (e.g. an MD5 checksum) which can contribute to votes of integrity (as explained in the case study of Paper D)
- SLA system characteristics
which are characteristics of the *General System Layer* in the BR in relation to a specific service level agreement, e.g. frequency of integrity checks across multiple pillars

Note that a SLA is based on specific pillars which represent a specific use of the BR, i.e. it is the SLA that defines a specific bit repository solution, or in Plato terms: a bit repository strategy.

For each type of characteristic there are categories of technical, organisational or economic characteristics. The different characteristics and their categories are illustrated in Figure 14.

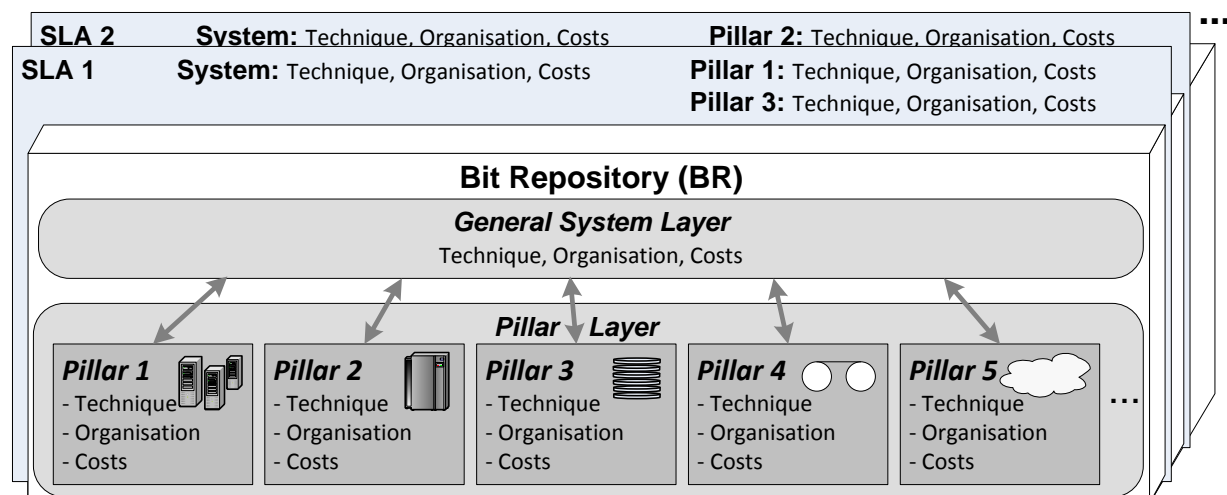


Figure 14 Characteristics for a bit repository with service level agreements in the general view

The illustrated media characteristics of the pillars in Figure 14 are only given as examples. There are other examples of media, e.g. microfilms. Note also that the media only represents some characteristics of a pillar, since other characteristics can be just as important, e.g. the geographical placement of the pillar.

Note also that in the general view there are no assumptions of where different replicas are placed, i.e. there can be several replicas placed on the same pillar.

There are several interrelations in the presented analysis, since e.g. SLA pillar points at a specific pillar. Furthermore results from calculations would be related to requirements for each of the SLAs. This called for a data model which contains characteristics and results, as well as a user interface to guide the different specifications and calculations. I chose Microsoft Access as basis for the development of the BR-ReMS, since Microsoft Access is well suited for quick specification of prototypes consisting of a data model and an interface, which can be based on data in the data model. The main form for entering the BR-ReMS, shown in Figure 15, was designed on basis of the analysis which pointed at the need to include definition and values for characteristics, specification of requirements along with specification of how values for the requirements are calculated, as well as means to perform the actual calculations and storage of the results.

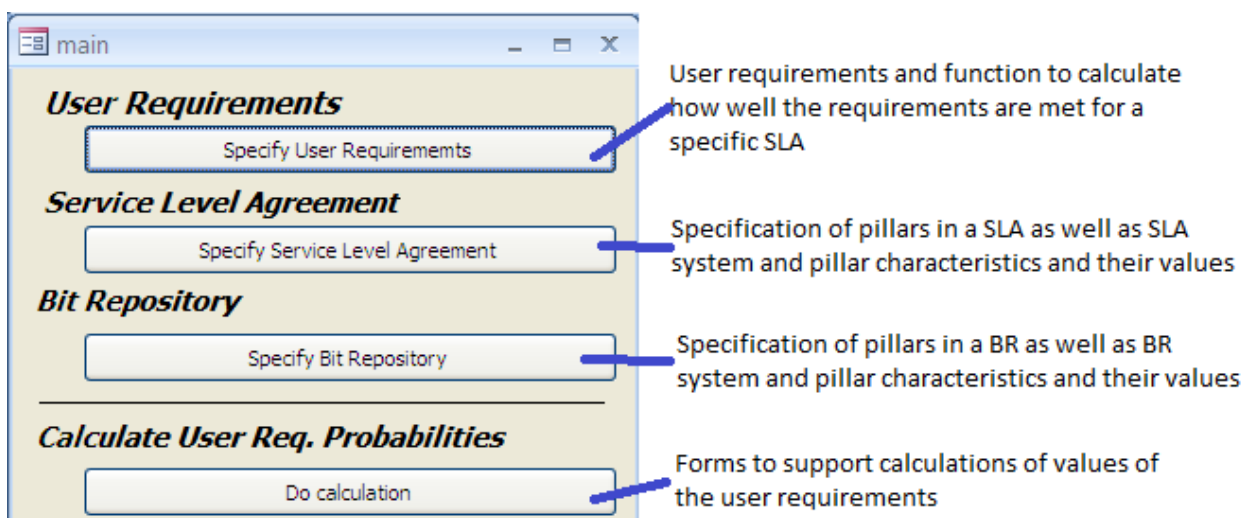


Figure 15 The BR-ReMS main form

In the following the data model will be briefly introduced, followed by an explanation of the different specifications and calculations. The user interface is based on the data model, and it provides a user friendly interface to add and modify data in the data model. In order not to make the description too confusing by mixing of the interface and the data model, the data model is only briefly described followed by explanation of some of the important parts of the user interface. The actual relation between the interface and data model is explained in footnotes. A full description of all forms in the interface as well as relations to the data model is provided in Appendix II "The BR-ReMS User interface".

The data model which the analysis led to is given in Figure 16. The data model is illustrated in an Entity-Relationship diagram using crowfoot notation. The entities are implemented in the BR-ReMS as physical tables. The simplifications consist in leaving out attributes such as 'sort order' solely used for presentations in forms in the user interface. In order to make the data model more comprehensible, I have also left out entities and references to entities with supplementary information like types of values of the characteristics, and categorisation of a characteristic in groups of technique, organisation, and economy. A full BR-ReMS data model diagram is provided in Appendix III "The BR-ReMS Data Model".

In the data model, it is the *Char* entity that contains definitions of characteristics. Definition of the BR, BR characteristics, and their values are defined in the entities contained in the blue rounded square named

BR, and likewise for the SLA. The dimmed BR entity is only included for presentation reasons, it does not actually exist in the database, since there are only one BR in this example. The names of the entities indicate what data they will include, where 'Sys' is short for system, 'Char' is short for characteristics and 'Val' is short for value. For example the *SlaSysCharVal* entity contains values of SLA system characteristics and the *BrPillarCharVal* contains values of BR pillar characteristics etc.

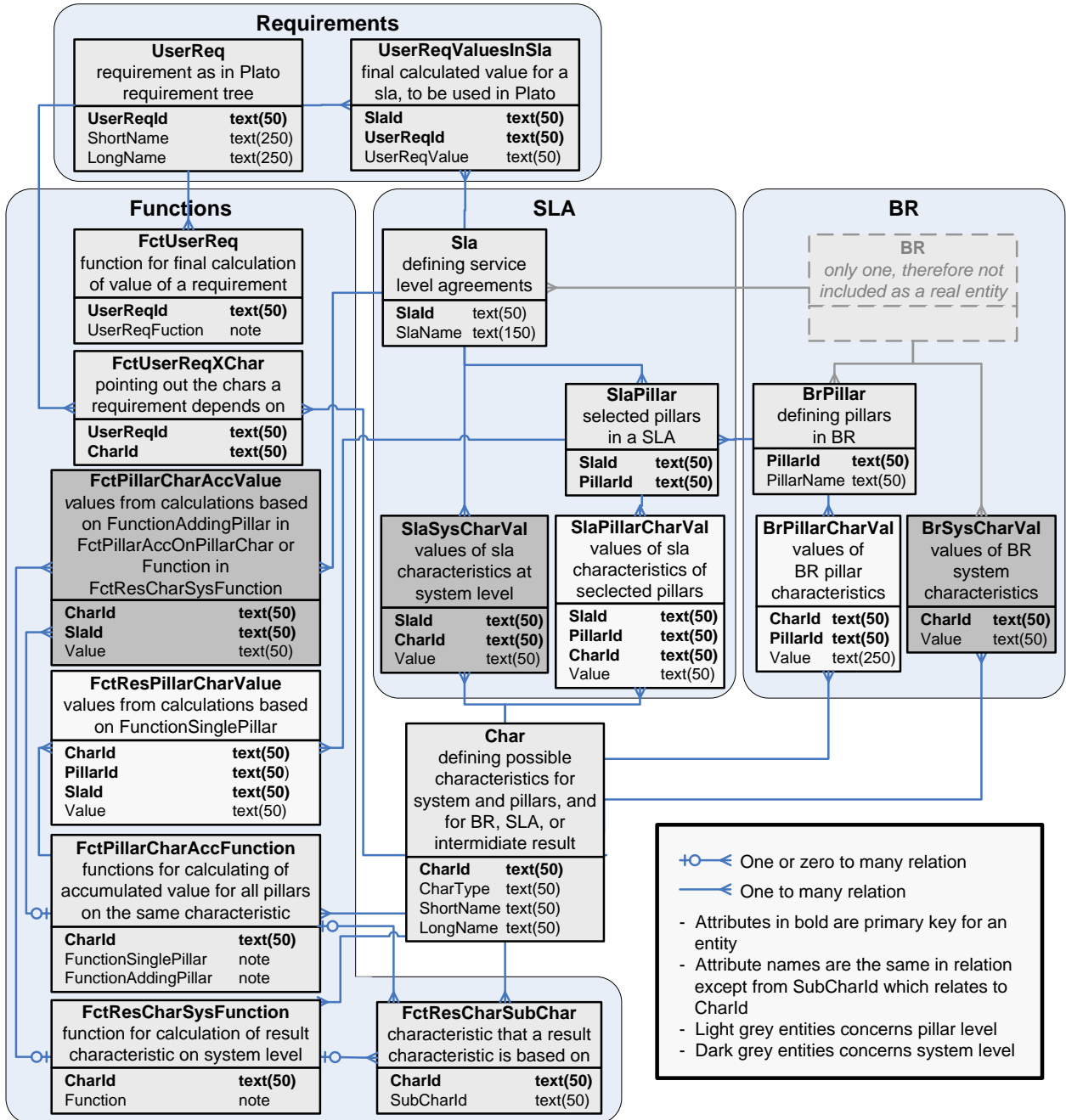


Figure 16 A simplified BR-ReMS data model of entities for BR and SLA information

Note that there are indicated to be more than four types of characteristics, which is explained later. The definitions of requirements are contained in entity *UserReq* and the final values from calculation to the individual requirements are contained in entity *UserReqValuesInSla*. The rest of the entities are related to the functions which are explained later as well.

As an example the forms for specification of characteristics and their values, Figure 17 shows the forms for the BR specification, activated by the *Specify Bit Repository* command button in Figure 15. The blue lines show which command button activates other forms.

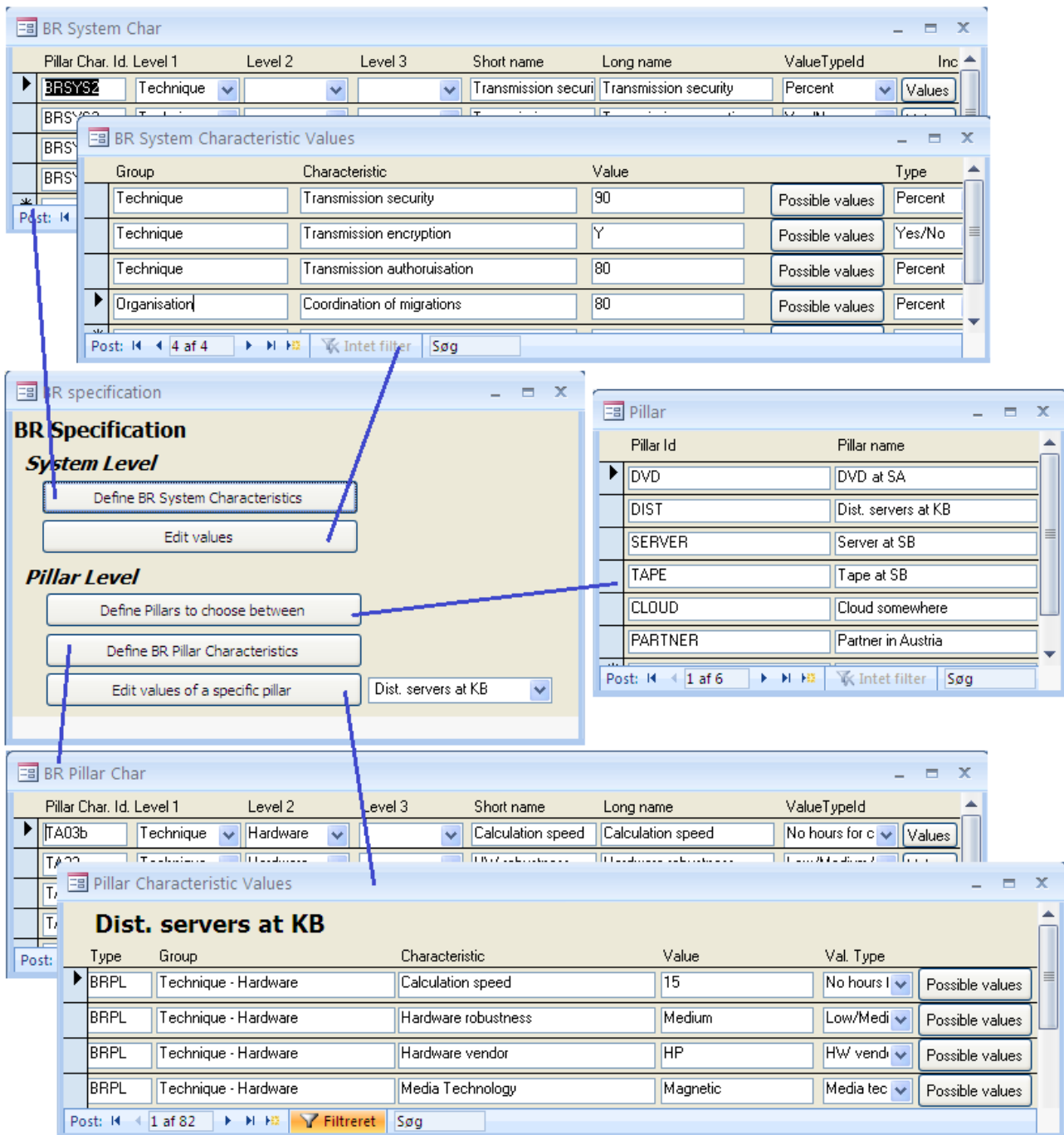


Figure 17 Forms for specification of a BR

The forms for SLA are implemented in the same way, and activated by the *Specify Service Level Agreement* command button in Figure 15. The form for the actual specification of possible user requirements is also similar, and activated via the *Specify User Requirements* command button in Figure 15. These forms for SLA and requirements specification can be found in Appendix II.

The definitions of the function for the calculations are more complex and belong to the specification user requirements. First of all, it must be noted that the functions are not programmed, but contain a pseudo-

description of what the function must do. That means the calculation of functions is done manually based on the function description and the arguments for the function. This choice was made because programming of the function would be very time consuming, and it would not add value to the goal of examining the evaluation methodology.

One challenge in specification of a function is that the computation can become very complex, if all computation is specified within one function. In the BR-ReMS, it is therefore allowed to break down the computation into computation of intermediate results. This means the function specification can use divide and conquer principles²⁰.

The intermediate results can be viewed as characteristics. The only difference from other characteristics is that these new characteristics require some extra computation. This is why the data model indicated that there were more types of characteristics than the BR and SLA characteristics, namely: the *result characteristics*²¹ which is the name for characteristics that have values that are results of a function. As for the BR and SLA characteristics there are results characteristics on the pillar and on the system level. Viewing them as characteristics allows the specification of functions to depend solely on function description and parameters in the form of characteristics.

An example of a system result characteristic is '*possible destruction by political attack*'. Figure 18 illustrates the form²² where the function to calculate the system result characteristic is specified.

The screenshot shows a window titled "Accumulated pillar result in separate result characteristic". Inside, there's a section "Function for result characteristic on System level" with a dropdown menu set to "Possible distruction by political attack". Below this are two sections: "Involved system Characteristica" and "Involved pillar Characteristica". Each section has a list of characteristics (e.g., "Min political distance", "Average political distance" for system; "No of full repl" for pillar) and buttons to "Add System Char", "Remove Sys Char", "Add Pillar Char", and "Remove Pillar Char". At the bottom, a text area contains the function logic:

```

IF <No of full repl> = 1 THEN ThisValue:=0,95
ELSEIF <Avarage political distance> < 5 THEN ThisValue:=0,95
ELSEIF <Avarage political distance> < 30 THEN ThisValue:=0,75
ELSEIF <Min political distance> > 150 THEN ThisValue:=0,75
ELSEIF [<Avarage political distance> > 1000] AND <No of full repl> > 3 THEN ThisValue:=0,05
ELSEIF [<Avarage political distance> > 150] THEN ThisValue:=0,25
ELSE ThisValue:=0,50

```

Figure 18 Function for a system level result characteristic

²⁰ Divide and conquer principles are here the original understanding of 'divide and conquer', not to be confused with divide and conquer algorithms which are based on recursion.

²¹ Intermediate results are introduced by allowing types of *Char* entity to include result characteristics on pillar and system level. Entities to contain values for intermediate results are entity *FctResPillarCharValue* for pillar values and *FctPillarCharAccValue* for system values.

²² This form is based on the entity *FctResCharSysFunction* for a given characteristic identified by a *CharId*.

In the example given in Figure 18, the specification of the calculation is given in the field *Function to calculate* 'Possible distrust by political att'²³. The arguments²⁴ of this function are specified by *involved pillar* and *system characteristics*, i.e. the characteristics are parameters to the function. Note that the calculation in this example is rather simplistic. In a real implementation it could instead be based on Bayesian statistics which can include assumptions of distribution of events [29].

Another challenge in specification of a function is that there will be varying pillars included for the different SLAs. This challenge is how to make the calculation independent of the specific pillars in SLAs. The values must somehow be referenced without knowledge of specific pillars in the BR, and which are within the given SLA. In the BR-ReMS this challenge is met by allowing specification of accumulated values from the underlying pillars. That means that pillar characteristics can have accumulated values²⁵ as well as values²⁶ for individual pillars. Figure 19 illustrates the form²⁷ for specification of a function to calculate values of pillar characteristics, as well as a function to calculate to accumulate pillars values.

Figure 19 Functions for a pillar result characteristic

Figure 19 provides the example of function specifications for the pillar result characteristics 'number of full replicas'. The argument²⁸ for the single pillar function is specified under *Involved pillar characteristics*. In the example, the argument is the SLA pillar characteristic 'object type', and the single pillar function is specified in the field *single pillar*²⁹. The accumulation function is given in the field *Accumulating pillars*³⁰, the arguments for these functions are not specified, since they are the pillar values of the given characteristic, i.e. in this example the 'number of full replicas' characteristic.

²³ Attribute *Function* in entity *FctResCharSysFunction*.

²⁴ Argument characteristics are represented in the *SubChar* attribute of the related *FctResCharSubChar* entity.

²⁵ All accumulated values are kept in the *FctPillarCharAccValue* entity.

²⁶ The individual pillar values are kept in the light grey entities, e.g. values for pillar result characteristics are contained in the *FctResPillarCharValue* entity.

²⁷ This form is based on the entity *FctPillarCharAccFunction* for a given characteristic identified by a *CharId*.

²⁸ Argument characteristics is represented in the *SubChar* attribute of the related *FctResCharSubChar* entity.

²⁹ Specification is in the *FunctionSinglePillar* attribute of the *FctPillarCharAccFunction* entity.

³⁰ Specification is in the *FunctionAddingPillar* attribute of the *FctPillarCharAccFunction* entity.

The form shown in Figure 19 is the same form³¹ used for specification of accumulating pillar values for BR and SLA pillar characteristics, but these would *not* include specification of function for single pillar values, since these values already exist.

Specifications of the functions for user requirements are specified in the form illustrated in Figure 20, i.e. function for the final calculations of how well requirements have been met. Here the function arguments³² are specified as either *involved pillar* or *system characteristics*. The function³³ itself is specified in field *General function involving above characteristics*.

Figure 20 The form for specification of calculations of user requirements

Note that the specified calculations are simple calculations resulting in the very low granularity scale with probabilities of fulfillment of the requirement given by the values “LOW”, “MEDIUM” and “HIGH”.

Turning to the manual calculation, the BR-ReMS offers an interface where the values can be entered via the *Do Calculations* command button in Figure 15. These calculation forms also provide access to method description in the form of the function descriptions, as well as values of characteristics that can be parameters to the function. As an example the Figure 21 illustrates the calculation form for calculation of the pillar result characteristics. The calculation is activated by the illustrated form *Do calculations to a Service Level Agreement*. The blue lines show which command button activates other forms. The illustrated form *Values from Pillar Functions for SLA* enables selection of the SLA pillar for which the values must be specified. Finally the form *Values for a Pillar in SLA* enables registration of the values calculated for the pillar characteristics. The red arrow shows where the values are entered. The red circle shows command buttons which will lead to forms with function description and values of other pillar characteristics which may be used by the function. Similar calculation forms are available for calculation of accumulated values, system level characteristics.

³¹ The form is based on the *FctUserReq* entity for a given user requirement identified by a *UserReqId*.

³² Arguments are contained in the *FctUserReqXChar* entity.

³³ The function description is contained in the *FctUserReq* entity.

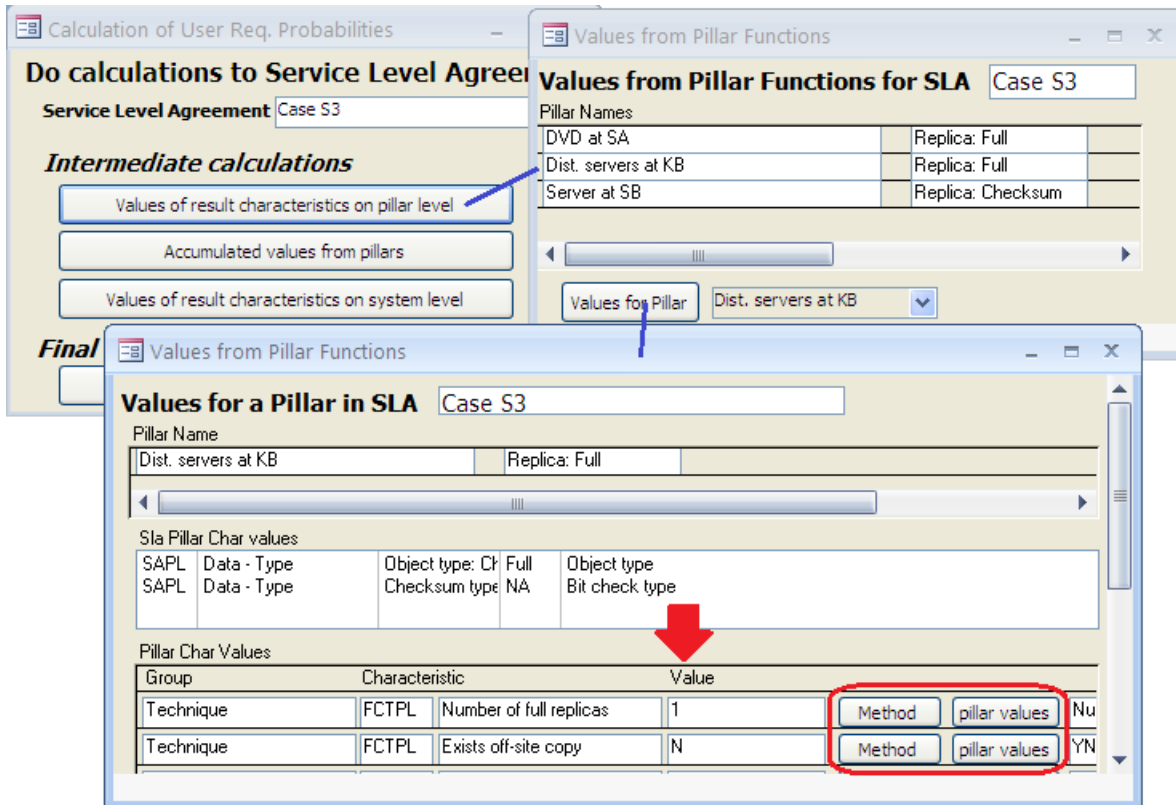


Figure 21 Entering pillar results for pillar level result characteristics

The BR-ReMS prototype additionally includes queries on reports on tables, e.g. reports with overviews of characteristics and functions. However, the basic information comes from tables implementing the entities presented in this section.

Note also that the way the BR-ReMS defines and structures functions probably could be done differently and improved. For an actual implementation it could also benefit from the use of other programming tools, and it should most definitely be based on a more stable database than Microsoft Access. The possibility of breaking down complexity into smaller intelligible units has been essential for this specific solution. However, when costs are included, this might instead consist of some sort of plugin of a cost model, which can directly give the needed intermediate results based on parameters calculated on the basis of the characteristics.

Although the BR-ReMS could be improved, it serves its goal as a prototype to investigate the feasibility of the evaluation methodology.

4.4. Use of the evaluation methodology

This section describes how the methodology can be used at different stages of the digital material life cycle by both providers and users of a bit repository. This section will include examples of how users or providers of bit repositories can use the methodology. It will also include a description of how it can contribute to auditing of a bit repository. Finally, it will include a description of how a SLA can be defined between the provider and user representatives of a bit repository.

4.4.1. Examples of bit repositories that can be supported by the evaluation methodology

As the methodology is based on the general view of a bit repository, there should be no limits to what kind of bit repository for which the methodology can be used. However, the BR-ReMS prototype will not fit any bit repository, as the implementations are different in respect to which types of characteristics are relevant as the basis for functions, and consequently, the function will be different as well.

One example of another bit repository other than the DK-BR is the LOCKSS system, which could be relevant for Kopal as well as members in Private LOCKSS Networks initiated by the MetaArchive. Since the LOCKSS system is a peer-to-peer system, there will be other relevant characteristics than for the DK-BR which forms the basis of the BR-ReMS prototype. That means there is a need for a BR-ReMS that is more specific to a peer-to-peer system. One example of differences is that the system layer is thinner in a peer-to-peer system, in the sense that much of the coordination and functionality is left to the peers. Other examples of characteristics can be found in the paper “2 P2P or Not 2 P2P” [148]. Other threats may also be specifically relevant to prevent for a peer-to-peer system, for example the threats described in the paper “Stealth modification versus nuisance attacks in the LOCKSS peer-to-peer digital preservation system” [149]. Thus the functions calculating probability for prevention of such threats are different for LOCKSS than for the DK-BR.

Another example is DuraCloud, which is more similar to the DK-BR. However, as noted in Paper E, there will need to be some knowledge of the cloud, in order to be able to specify any of the pillar characteristics. For instance, if a replica in the cloud can be placed anywhere, it will be hard to say whether two replicas are independent from a geographical viewpoint.

The general view of a bit repository could suggest that there is only one general system layer that handles communication with all pillars in the same way. However, this is not the case. One example where there are differentiated treatments of pillars can be found for private pillars like scenario B illustrated in Figure 9 in chapter 2. The BR-ReMS does not currently cover the possibility of specifying differentiated treatment of pillars in the general system layers. However, this could be done by enhancing the BR-ReMS. Note that the methodology would be the same, if such enhancements are made. Furthermore, such an enhancement could also cover cases based on LOCKSS where an additional off-line replica is wanted, since the LOCKSS system is only suited for on-line replicas [157].

4.4.2. Different stages where the methodology can be used by BR users

From a BR user’s perspective, the methodology can be used at different stages of the life cycle for digital material to be bit preserved. The most obvious stage is when evaluating bit preservation solutions for known bit preservation representation, as it is the case in the evaluation example in Paper E. Other stages can be in evaluation of choice of bit preservation representation the material, or in re-evaluation of bit preservation strategies, if conditions for the bit preservation have changed.

There are different examples where the methodologies can be relevant as part of evaluating the choice of bit preservation representations. One example is the choice between different representations coming from a digitisation process, as described in the case study in Paper B. However, there will be a need for supplementary methodology in order to include the initial digitisation costs in the total evaluation of costs. Another example is different choices of preservation formats which were the case for the Royal library of the Netherlands and the Bavarian State Library. Here an evaluation can assist in evaluating the choice based on the differences in storage consumption and required bit safety for the two formats JPEG2000 and TIFF. However, this evaluation cannot be supported on the basis of the current BR-ReMS

prototype, since it does not yet include economy and parameters of required storage. Furthermore, such an evaluation would need refinement of the bit safety requirements expressing differences between acceptable losses of bits in bit-streams.

There will inevitable be changes in conditions for bit preservation for any bit preservation solution, since technological evolution will mean changes of media over time. This might for instance be an upgrade of an operating system, or a media migration where the characteristics change on e.g. costs, processing power, or longevity of media. If changed characteristics influence the calculations of how requirements are met, then a re-evaluation will be needed in order to see if the changes are acceptable. With a production version of a BR-ReMS, such a re-evaluation can be limited by use of the BR-ReMS. The changes will result in an update of the relevant characteristics and a re-calculation of fulfilment of requirements can be performed.

4.4.3. Use of the methodology by BR providers

There are two ways that the evaluation methodology can be of benefit for a BR provider. Firstly, it can be the case that the BR provider wants to expand the BR with changed services that can be offered to BR users. Secondly, it can be used to contribute to clearer definition of when SLAs must be re-negotiated, if conditions in the BR change. In the DK-BR and in a Private LOCKSS Network, the BR providers are the ones providing pillars and the overall administration of the BR.

Examples of expansion of the BR with changed services can for instance be possibilities for mass processing on a specific pillar, or adding new pillars to the BR. Adding mass processing can give extra availability services, but also added confidentiality issues for the specific pillar. Adding a pillar can broaden the scope of possible independence of pillars, e.g. by adding a pillar in a foreign country or based on microfilms. A foreign pillar may however only be suitable for materials that can be placed under foreign legislation. A microfilm pillar can also give additional solutions for confidential material, but the data are not very accessible and thus it entails challenges for function preservation requirements.

A SLA should always be clear on when a service level is actually delivered. This can be hard to formulate in terms of when the requirements are fulfilled unless they are defined in terms of how the fulfilment is measured. The evaluation methodology can here give a clearer basis for measuring, when a change e.g. in a pillar will affect SLAs using the pillar, and thus must trigger a re-negotiation of the SLAs. It can mean that there will be fewer cases where BR-users must be given notice of changes and consequences of changes, since cases in doubt can be settled by internal use of the BR-ReMS and re-evaluation.

4.4.4. Continuous audit and evaluations of bit repository solutions

In order to ensure that bit preservation solutions are sustainable, there will need to be some sort of regular audit of whether the bit repository providing solutions is trustworthy, and that a chosen solution is the best suited solution.

The evaluation methodology can only provide means to evaluate and choose a solution that is best suited for specific requirements, but it cannot contribute to trust in whether the given characteristics are properly valued, e.g. whether operation procedures exist and are followed.

Although TRAC and DRAMBORA are not strong on storage aspects, there are many important aspects of a bit repository which they cover. Concerning bit safety expressed as how well risks of losing bits are prevented, DRAMBORA specifically addresses risk management through prevention and detection. Auditing of how well risk management is implemented is an important aspect that is only covered

indirectly in the evaluation methodology in function definitions, which would have to be based on characteristics of existing risk management. It is only regular audits of the actual implementation of risk management that can contribute to the trust of the risk management.

Specifically for trust in specific characteristics, this can also be obtained through use of iRODS, where a simple example would be to set up rules to check whether values of specific characteristics like ‘frequency of integrity check’ are the actual values. In this sense iRODS can contribute to trust on a day-to-day basis. It should however be noted that there can be overhead of implementing iRODS, and in the example of integrity checks, it may not be feasible to implement integrity checks as a micro service on large scale systems with billions of files, thus such checks may be better implemented directly as part of a bit repository.

There is one aspect where the evaluation methodology can contribute to audit of a specific bit preservation solution. This is the aspect of whether conditions have change in a way that affects the chosen solution or the requirements for the solution. In such a case the evaluation methodology can be used in re-evaluations of whether the choices made are still the best choice for the specific requirements.

4.4.5. Service level agreements

The concept of a SLA is an important part of the methodology, since it is this SLA that defines specific bit preservation solution. There exist many varying definitions of SLAs, and related agreements [193]. This thesis will not go into details about these definitions, but only refer to SLAs in the definition given earlier. Based on that definition, here follows a discussion of some of the aspects that should be considered for a SLA.

Even with the very narrow definition of a SLA, one of the points in Paper E is that changes in the SLA can have an impact on whether the SLA can fulfil requirements. A simple example is that a change of bit integrity check frequency can affect requirements related to bit safety. An example that involves changes in the pillar is the frequency of checksum calculation. A major change in the SLA would be changes in choices of involved pillars.

There are still a lot of issues to be investigated about SLA concerning what it should include, who the actual parties are who agree to a SLA, how to control that the solution fulfils the SLA, and what the consequences should be if the SLA is not fulfilled. More inspiration can e.g. be found in the paper “Service Level Agreement (SLA) in Utility Computing Systems” [193].

As mentioned in Paper E, there can be several organisations involved in operating different parts of the BR. Therefore one of the outstanding issues is to find out who the parties to a SLA contract should be. Examples of scenarios could be that representatives from all concerned pillars as well as the responsible party for the general system layer should be involved. Another scenario could be that only one organisation responsible for the bit repository should be involved.

In order to control whether a SLA is fulfilled, it will need to contain measurable criteria for when it is fulfilled. As mentioned, the calculation function from the BR-ReMS can assist in setting such measures, but it will have to be accompanied by audits and continuous evaluation on basis of reporting that is explicitly stated as required in the SLA as well.

Definition of calculation functions of the BR-ReMS will raise issues. There may be issues related to how they contribute to choice of a solution, and to control of fulfilment of a SLA. They may be other issues

related to choice of persons to define the selected characteristics. The reason why these issues are inevitable is that there will be different opinions on how different fulfilment of requirements should be calculated. Even though the BR-ReMS offers documentation of the calculations, procedures are needed for when, and by whom, these calculations can be maintained or altered. Changes of calculations can mean a change in whether a requirement is fulfilled or not. Thus, if a SLA is based on fulfilment of requirements, a change made after a SLA is agreed upon can mean that a SLA is no longer fulfilled.

Another issue is whether definition of calculations can be reviewed adequate by the users of the BR. The calculations may be quite complex and hard to understand. Therefore users of the BR may need a third party to assist in evaluation. A third part can also be used to eliminate risk that calculation accidentally has been designed to fit specific purposes or needs by either the BR representative or the user of the BR.

As mentioned in Paper D sanctions or escalation procedures are needed in order to enforce that the requirements specified in a SLA are met. This is especially important concerning bit safety and confidentiality, since loss of data or leak of confidential material cannot be undone. Sanctions or escalation procedures will therefore be needed before such violations occur, if possible. An example could be to specify what the notification should be for media migration. Furthermore it could include a list of possibilities for users to postpone media migration, e.g. if internal replicas are at risk at the same time.

A SLA cannot solve everything. No matter how precise it is, there will always be issues of how measures are calculated and a matter of trust in the organisation(s) taking care of the data. Trust can also be based on the motivation of fulfilling a SLA, for instance public and commercial organisations can have different motivations and different focuses on how they fulfil SLAs. Furthermore, if data is physically placed abroad, there may be legislation which can affect e.g. trust in confidentiality.

An example where a SLA cannot help is in the challenge of funding for bit preservation. Therefore bit preservation will always be in jeopardy, since nobody will be able to foresee funding. No methodology or SLA can assist solving this challenge. However, this will always be the case no matter which type of preservation is in question.

4.5. Summary

The methodology is the first of its kind to provide a way to evaluate the best choice between different bit preservation strategies in form of bit preservation solutions. It supports choice between bit preservation solutions based on evaluation of how well the solutions meet the requirements for bit preservation of specific digital material.

The methodology can be used for any bit preservation system which can be described in the general view of a bit repository, which was exemplified by LOCKSS based systems. The methodology can be used by providers of a bit repository as well as users choosing a bit preservation strategy for their digital material.

In a holistic approach to bit preservation, an optimal solution must fulfil various requirements for the bit preserved material, and requirements for bit preservation taking as many aspects from the “whole” into account as possible. The solution must be continuously evaluated in order to ensure that the optimal solution is chosen if conditions or requirements for the bit preserved material change.

5. A Simple Example Using the Results

This chapter is aimed at readers with a technical background. It provides simple examples to illustrate how the different results of this thesis can be used for digital material that is to be bit preserved.

In order to keep the example simple, limitations will be made to the scope of aspects of the digital material included. For instance, the example will only consider some structural metadata, although other metadata would have to be taken into account in the general case. It will be described using the terms defined in the IR-BR model described in Paper D, where digital material is ingested to an institution repository (IR) which leaves bit preservations to be dealt with in the bit repository (BR).

The digital material considered is a simplified book consisting of two pages, which must be made available to the public as in the Archive for Danish Literature described in Paper A and Paper B. The origin of the book is an analogue copy that is digitised.

It is assumed that, in order to optimise preservation costs, a prior analysis of the digitisation process has been carried out. This includes evaluation of possible bit preservation solutions and their ongoing costs for different results of different digitisation processes with different initial costs.

5.1. IR-BR Ingest and Representations of Digitised material

In this example the choice is a digitisation process where each page is scanned and given in a TIFF file, and characters from the pages are extracted using an Optical Character Recognition (OCR) program like e.g. FineReader. The resulting OCR text is manually encoded using the TEI-P4 guidelines for text encodings in XML [170]. The result of the digitisation is given in the following files:

- 1 TIFF file with image of page 1
- 1 TIFF file with image of page 2
- 1 TEI-P4 file with OCR & encoded text of page 1 and 2

Figure 22 illustrates the content of these files.

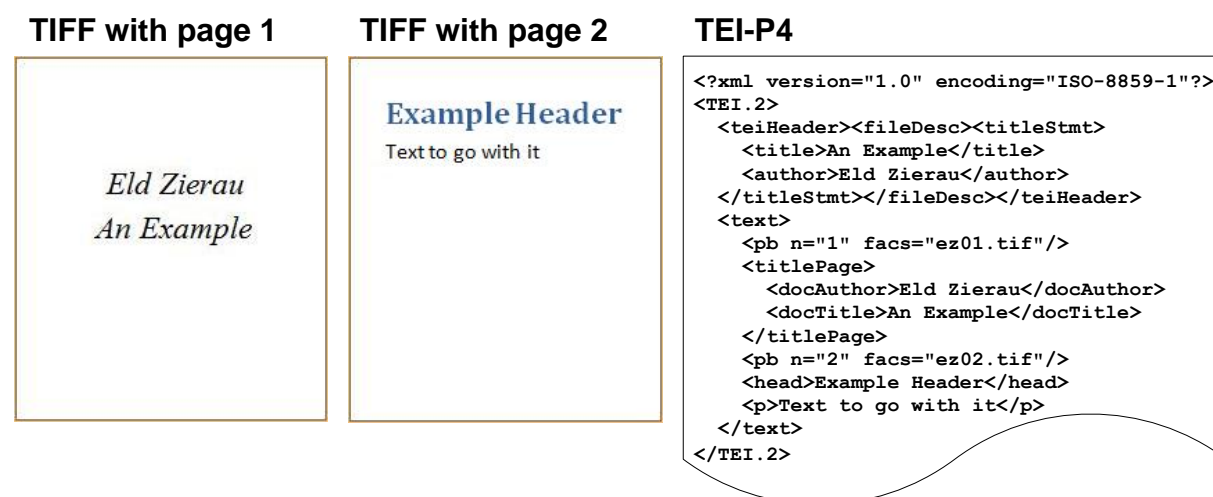


Figure 22 Contents of files from digitisation

Looking at the ingest flow, the IR-SIP (Institution - Submission Information Package), which is ingested into the IR, is a package consisting of the results from the digitisation, e.g. delivered in a ZIP file. It is assumed

that there are techniques to ensure that the contents of the ZIP are received correctly. Figure 23 illustrates the ingest of the digitisation result to the IR.

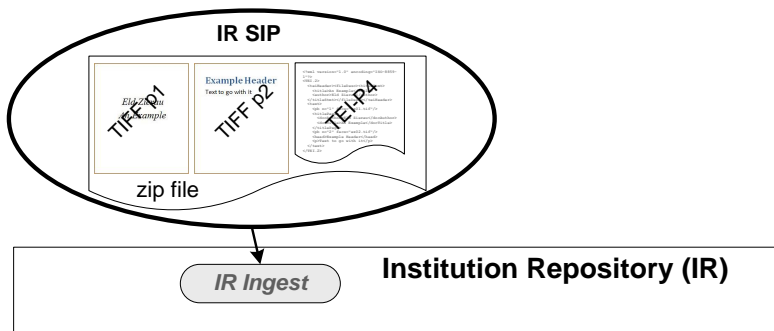


Figure 23 Ingest in IR

In this example it is assumed that it is the IR that assigns opaque universally unique identifiers (UUID) to the different elements, in order to ensure future reference. It is also assumed that the quality of the encodings is validated as part of the ingest process. This could for instance include a check of a sample of the OCR text. The check could be done by trimming the contents of the TEI-P4 file by removing all encodings, line breaks and extra spaces, and then matching it manually by inspecting the TIFF files. Acceptable error rates, e.g. 1% will have to be part of the validation procedures.

The decisions on the final preservation representations of the book are assumed to be based on prior analysis of how it must be disseminated using the representation concept described in Paper A. It is also assumed that this analysis has contained modelling considerations supporting a migration strategy for functional preservation of the material, as described in Paper C. The representations in this example are illustrated in Figure 24.

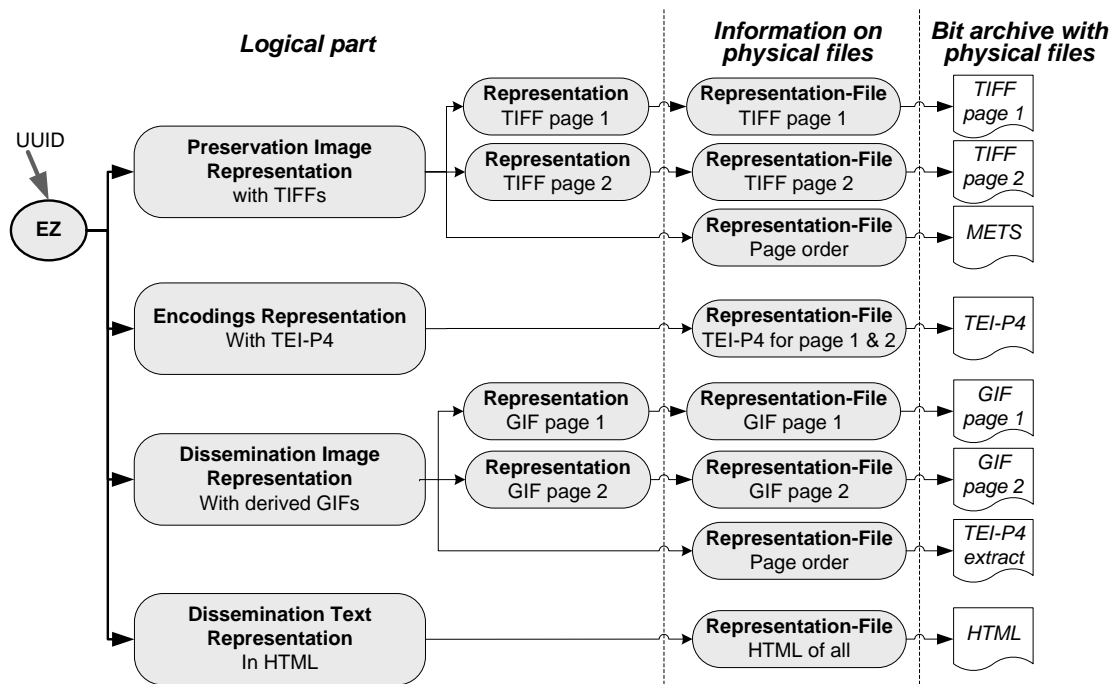


Figure 24 Simple representations for preservation and dissemination

Focussing on the dissemination representations, it was the ingested TEI-P4 file that included the structural information for the *Dissemination Image Representation*. In this example, the structural information is

extracted for the image representation, which includes GIF files derived from the ingested TIFF files. The choice of GIF-files for dissemination is due to the fact that they are less storage consuming than the TIFF files. The *Dissemination Text Representation* is HTML derived on basis of the ingested TEI-P4 file in a way, which can present the encodings in a web browser.

The way that the dissemination representations are derived from the preservation representation is information that is attached to the relations between the representations, e.g. there is a *derived from* relation from the GIFs to the TIFFs which includes description of how the individual GIF is derived from a TIFF using the ImageMagick tool³⁴. Finally the HTML file is derived from the TEI-P4 file.

The contents of the GIF files are similar to the TIFF files, the derived XML files for dissemination representations are given in Figure 25.

HTML with all

```
<html>
  <body>
    <p></p>
    <table width=400>
      <tr><td><center>
        <p><b><i>An Example</i></b></p>
      </center></td></tr>
      <tr><td>
        <center>
          <p><b><i>Eld Zierau</i></b></p>
        </center>
      </td></tr>
    </table>
    <p></p>
    <table>
      <tr><td><p><b>Example Header</b></p></td></tr>
      <tr><td><p>Text to go with it</p></td></tr>
    </table>
  </body>
</html>
```

TEI-P4 extract

```
<?xml version="1.0"?>
<TEI.2>
  <pb n="1" facs="ez01.tif"/>
  <pb n="2" facs="ez02.tif"/>
</TEI.2>
```

Figure 25 Simple representations for preservation and dissemination

Turning to the preservation representation, the metadata for the representation is chosen to be expressed in the METS standard format using the METS XML schema [101]. The structural metadata information of the preservation representation is included in METS file derived from the ingested TEI-P4 file, i.e. the METS file contains description of the order of the TIFF files.

Each physical file to be bit preserved also has a related METS file, with the various metadata of the files, e.g. technical metadata for later functional preservation actions. This is *left out* of the illustrated representation in Figure 24 in order to keep it clear, but all³⁵ the METS files will be explained in the following. Selected contents of the derived METS files for the preservation representations are given in Figure 26. The figure includes the METS file illustrated in Figure 24 with page orders as well as the three METS files for the individual physical files for preservation representations. In real examples, the METS files would contain much more metadata than the metadata included in Figure 26, e.g. technical

³⁴ ImageMagick is a software suite to convert and write images in a variety of formats. Further information can be found on <http://www.imagemagick.org/script/index.php>.

³⁵ In a real case all the METS information for the *Preservation Image Representation* could be placed in one METS file, but this is not considered here, since it would complicate the presentation of the example.

metadata for the TIFF files would be included in the METS file for the individual TIFF files. There are given examples of the mentioned PREMIS fields in the METS on the full representation. PREMIS parts are written in purple, while METS parts are written in green. Note that it is *not* all PREMIS preservation level fields that are given, and that METS files for the images would include such information in a real case. All values of PREMIS fields and identifiers are highlighted with increased font size and blue colour. The choice of universally unique identifiers is UUID specified in an URN. The identifiers are specified as URIs on form of an UUID in an URN on form `urn:<UUID>` [87], but the UUID part in the example is replaced with more readable text.

METS on full TIFF representation (some structure & some PREMIS)

```

<mets OBJID="urn:uuid:ezTIFFSmdUUID">
  <amdSec>
    <techMD CREATED="2011-08-01T12:10:00">
      <mdWrap MDTYPE="PREMIS:OBJECT">
        <xmlData>
          <object xsi:type="representation">
            <linkingIntellectualEntityIdentifier>
              <linkingIntellectualEntityIdentifierType>
                URN
              </linkingIntellectualEntityIdentifierType>
              <linkingIntellectualEntityIdentifierValue>
                urn:uuid:LOGICEzUUID
              </linkingIntellectualEntityIdentifierValue>
            </linkingIntellectualEntityIdentifier>
          </object>
        </xmlData>
      </mdWrap>
    </techMD>
    <digiprovMD CREATED="2011-08-01T12:10:00">
      <mdWrap MDTYPE="PREMIS">
        <xmlData>
          <preservationLevel>
            <preservationLevelValue>
              HighBitSafety
            </preservationLevelValue>
          </preservationLevel>
        </xmlData>
      </mdWrap>
    </digiprovMD>
  </amdSec>
  <structMap>
    <div>
      <mptr ID="urn:uuid:ez01mdUUID" LOCTYPE="URN"/>
    </div>
    <div>
      <mptr ID="urn:uuid:ez02mdUUID" LOCTYPE="URN"/>
    </div>
  </structMap>
</mets>

```

METS TIFF image 2

METS TIFF image 1

```

<mets OBJID="urn:uuid:ez02mdUUID">
  <fileSec>
    <fileGrp>
      <file ID="urn:uuid:ez01srcUUID">
        <Flocat
          LOCTYPE="URN"
          xlink:href="urn:uuid:ez01srcUUID" >
        </fileGrp>
      </file>
    </fileSec>
    <structMap>
      <div>
        <fptr FILEID="urn:uuid:ez01srcUUID"/>
      </div>
    </structMap>
  </mets>

```

METS TEI-P4

```

<mets OBJID="urn:uuid:ezTEImdUUID">
  <amdSec>
    ..
    <linkingIntellectualEntityIdentifierValue>
      urn:uuid:LOGICEzUUID
    </linkingIntellectualEntityIdentifierValue>
    ..
  </amdMD>
  <fileSec>
    <fileGrp>
      <file ID="urn:uuid:ezTEIsrcUUID">
        <Flocat
          LOCTYPE="URN"
          xlink:href="urn:uuid:ezTEIsrcUUID" >
        </fileGrp>
      </file>
    </fileSec>
    <structMap>
      <div>
        <fptr FILEID="urn:uuid:ezTEIsrcUUID"/>
      </div>
    </structMap>
  </mets>

```

Figure 26 Contents of derived METS files

Note that the logical identifier specified in PREMIS is the same in METS on the full TIFF representation and file and the TEI-P4 representation, since they are representations of the same intellectual entity.

Looking at the entire ingest flow, it is the IR that prepares packages that are to be bit preserved in the BR. In order to secure identifiers to files, there will in this example be used WARC as the package format. In this example there will be made one WARC file for each pair of physical files with its metadata specified in METS and one for the METS file with structures.

Figure 27 illustrates the entire ingest flow of the digitisation result to the IR. The dashed circle represents parts for the IR AIP, since a real AIP would require much more information in order to have all information of the elements in the AIP. The packages generated by the IR for bit preservation are the BR SIPs which are ingested into the BR.

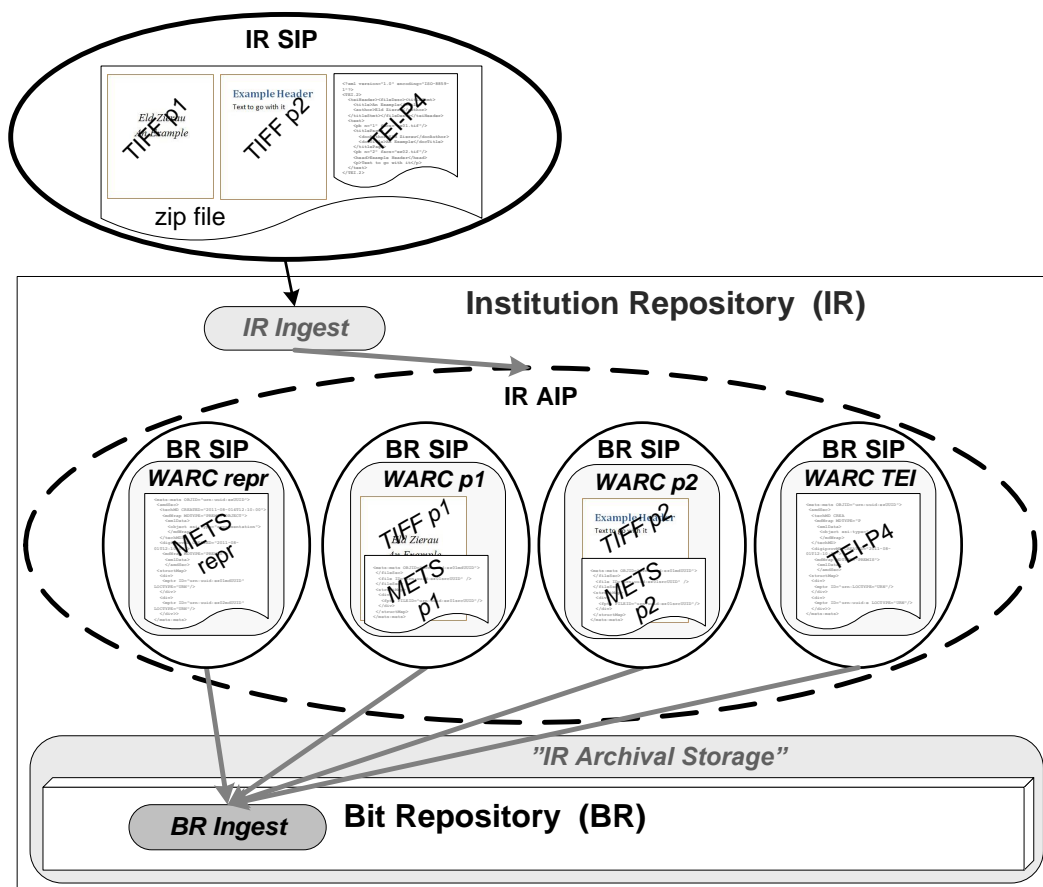


Figure 27 Ingest in IR and BR

For completeness Appendix IV provides an example of the *WARC repr* file and for the *WARC p1* file based on the METS given in Figure 26.

The derived GIFs and HTML files are not bit preserved in this example. Note that if fast access to the GIF files is required at all times then the derived GIF files could need to be ingested into the BR as well (with low bit safety). This could be needed to fulfill access time requirements in case GIFs are lost. The need will be for cases where computation of derived GIFs along with the BR access time to the originals cannot meet the requirements for fast dissemination access.

It is assumed that the different bit preservation requirements of the different representations are the following:

- The preservation representation needs medium bit safety, since the original analogue book still exists and can be rescanned. Requirements for access are also medium (e.g. average an hour, max one day) since we in this example will have a separate dissemination representation which in most cases will be available. Confidentiality is not an issue here, and costs are limited to a certain budget. Note that the representation is already based on consideration of later possible migrations.

- The *encodings representation* needs medium/high bit safety, since manual work is involved which in this hypothetical example is considered hard to reproduce. Confidentiality, availability, and costs are the same as for the preservation representation.
- All *dissemination representations* only need low bit safety, since they can easily be derived from the preserved representation. However the requirements for access are much higher since they must be speedily displayable. Confidentiality and costs are the same as for the other representations.

5.2. Evaluation of Bit Preservation Solutions

The requirements differ for the files belonging to different representations; therefore SLAs must be made for each set of requirements, which in this example corresponds to each of the representations. Note also that it is not necessary to use the same bit repository for the representations. For instance for the dissemination representations, which do not require high bit safety, it may be considered reasonable to find another bit repository solution.

This example only looks at elements in the SLA for the *encodings* representation, i.e. the TEI-P4 file. We will use the evaluation methodology described in Paper E, to see if a bit preservation solution will meet the requirements.

The requirements tree for Plato used by the evaluation methodology is sketched in Figure 28, but for simplicity, it only includes bit safety.

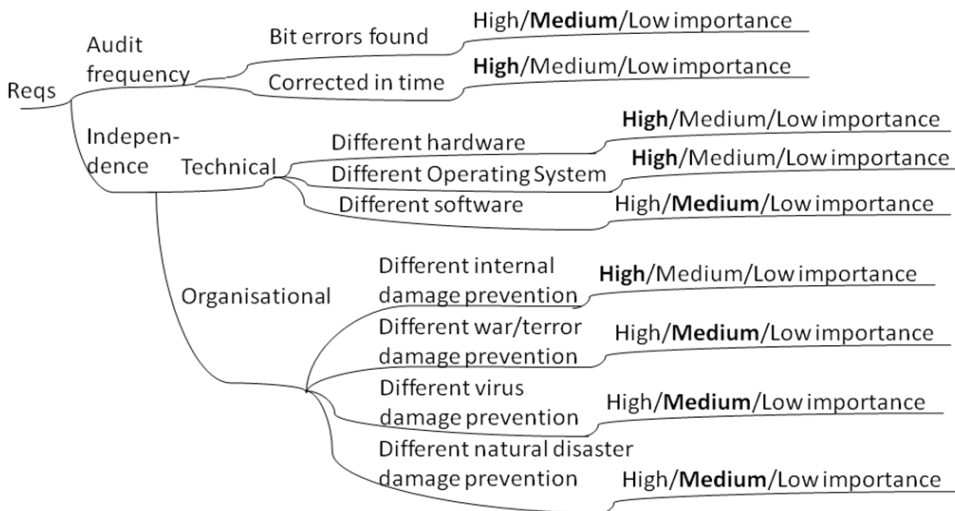


Figure 28 Requirements tree

The exact importance of fulfillment of the individual requirements for the TEI-P4 file is marked by the bold values in Figure 28. Before we can use the BR-ReMS to help us evaluate a bit repository solution, we first need to define the values for different characteristics for:

- *Implementation of bit repository*

This example only includes a description of one disk pillar placed in the organisation of the Royal Library of Denmark and one DVD pillar placed in the organisation of the Danish State Archives of Denmark. On the general system layer we assume that some sort of organisation operates a system that supports communication with the pillars, running checksums checks etc.

- *Use of bit repository*

In this example one full copy of data is placed on each of the pillars, and with a requirement of a bit integrity check every 4th month.

The elements in the evaluation methodology are illustrated in Figure 29 along with an example of specific values of characteristics.

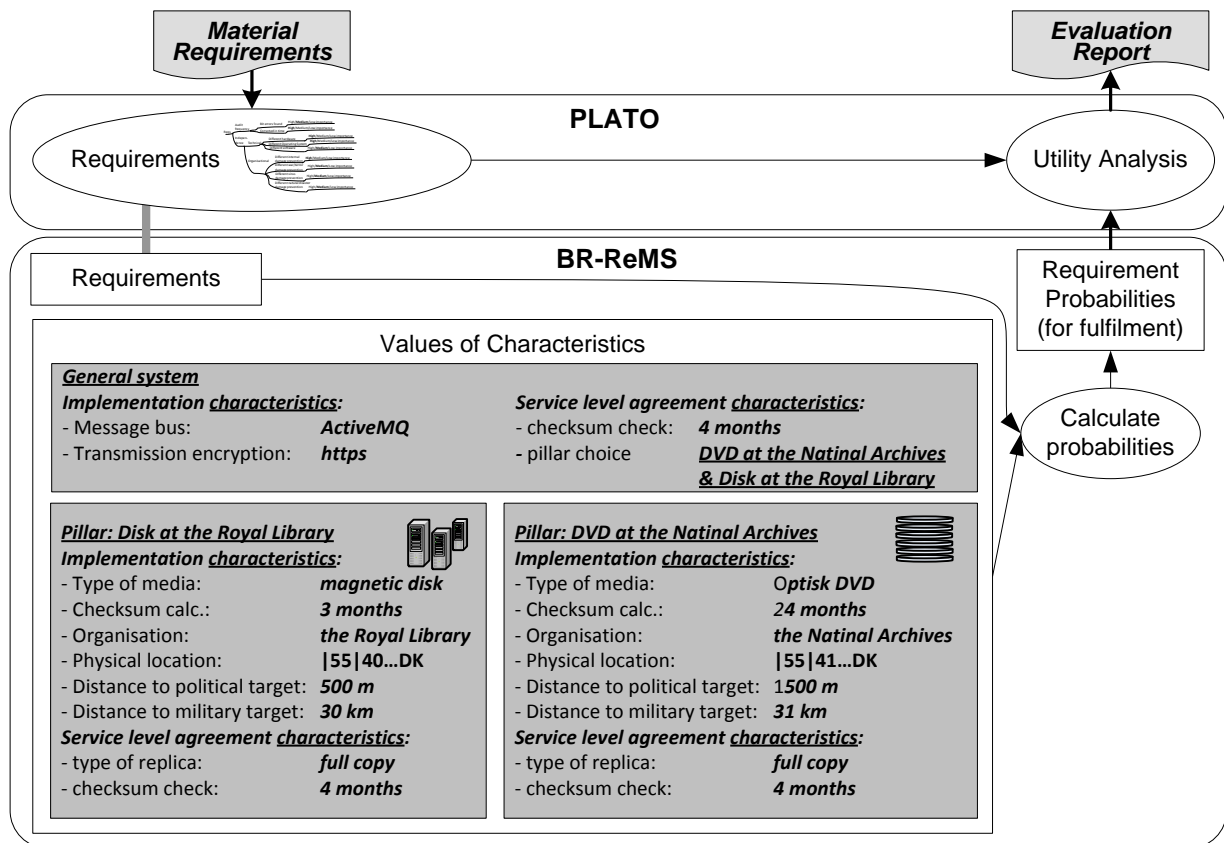


Figure 29 Example of input to evaluation methodology

Now calculations can be made based on the specified requirements and the given characteristics. We will only look at a simple example of a calculation function. This function calculates the probability of meeting the requirement from branch *independence - organisational*:

“difference in war/terror damage preventions”

In the current prototype of the BR-ReMS the calculation is done manually based on:

- Distances between full replicas (which is 1 km calculated on basis of the physical locations)
- Assessment of replicas
 - Assessment of replica on disk pillar at the Royal Library
 - Distance to political targets (500 m as given in Figure 29)
 - Distance to military targets (30 km as given in Figure 29)
 - Internal organisational risk mitigation procedures (Middle based on other values)
 - Assessment of replica on DVD pillar at the Danish State Archives
 - Distance to political targets (1500 m as given in Figure 29)
 - Distance to military targets (31 km as given in Figure 29)
 - Internal organisational risk mitigation procedures (High based on other values)

- Assessment of General System characteristics
No characteristics of interest in this simple example

The short distance between the pillars and the short distance to a political target will in this example result in a low probability to fulfill requirements on *difference in war/terror damage preventions*.

It is not likely that the bit preservation solution defined by the SLA will satisfy requirements for bit safety, since there are only two full replicas in this example. The reason is that a bit integrity check only will be able to point at errors, but it cannot point at a correct replica, since that would require at least three votes. Therefore a solution with more replicas must be considered in this example.

This example is too simple to finalise calculation of the requirements probabilities and subsequent utility analysis in Plato. Instead Appendix I provides full details of the calculation described in Paper E. The Appendix I furthermore provides more details on use of Plato.

6. Conclusions

The *holistic approach to bit preservation* has set the framework to focus on bit preservation, but also to take into account as many aspects related to bit preservation from the “whole” as possible. The aspects included in the holistic approach are only those that affect choice of an optimal bit preservation strategy, which can meet the all relevant requirements for bit preservation.

The three main results contributing to this *holistic approach* are the *IR-BR model* to separate the bit preservation tasks of digital preservation activities, the *representation concept* to support analysis of the final form of the bit preserved digital material, and the *evaluation methodology* to assist the selection of a bit preservation solution.

The *IR-BR model* contributes to delimit and define bit preservation in a repository which can be shared by the repositories of more than one institution. The model provides a framework to achieve a clear understanding of terminology, and as a basis to analyse placement of general repository functions in terms of OAIS functions in functional entities, i.e. what part of the OAIS functional entities must be covered by the IR and which must be covered by the BR.

The *IR-BR model* can generally be used to analyse separation of bit preservation, since it contributes to definition of separated bit preservation on a conceptual level. Furthermore, it gives a general basis for arguing that a system with a separated OAIS compliant bit repository can be OAIS compliant, including cases where the bit repository is placed in another organisation.

The *IR-BR model* does make sense within OAIS, which means that it can fill the gap of a missing model to describe separated bit preservation repositories. Additionally, it makes the IR-BR model even stronger that it is defined within an already well known reference model.

Some more concrete results from the *IR-BR model* are that, concerning the interface of the IR and the BR, the model points to the importance of audit trails and identification of bit streams for bit preservation. Furthermore, the case study points to the need to express requirements of information security for a BR, and investigation of a SLA in cases where the BR was a separate organisation.

Turning to the *representation concept*, this concept contributes to the *holistic approach to bit preservation* by providing a basis for analysis of the optimal preservation representation for bit preservation.

The *representation concept* is a concept that is new in the way it gives a nuanced perspective on representations that can be used as a basis for preservation activities and as a basis for dissemination. The strength of the representation concept is that it supports analysis of digital material that must support different purposes in preservation and dissemination.

The *representation concept* supports analysis of how the form of material can support requirements from preservation and dissemination, where the form is expressed in files with and information about structure and metadata for the material. In cases where there are different representations for different purposes, the concept supports analysis of relations between these representations and their possibly differing significant information. The issues, which can be crucial for files in an optimal representation, were e.g. the chosen file formats, the way the files are created, migrated or compressed. The crucial aspects of metadata are e.g. how the material becomes identifiable on a permanent basis, and the representation of

structure within a representation, which can be crucial for supporting expression of many-to-many relations.

The *representation concept* also assists in revealing different possible implementations of representations, which can require a different bit preservation solution. The important representation aspects here can both concern the formats of the files used in the representation, structure of relations between files in the representation, and information about the representation expressed in metadata.

Furthermore the *representation concept* assists in revealing the information security requirements that there may be for the representation to be bit preserved, where such requirements must be taken into account when formulating the requirements for the bit preservation solution for the representation.

In the *holistic approach to bit preservation*, the interesting part of the *representation concept* is the analysis of the representation for bit preservation, and the various requirements for the representation as well as the bit preservation of the representation.

The *evaluation methodology* contributes to the *holistic approach* by providing a methodology to assist decision making between different bit preservation strategies, which is a step towards choice of an optimal bit preservation solution.

The *evaluation methodology* fills a gap in the means to evaluate bit preservations strategies. One of the main challenges is that there are currently no suitable ways to define bit safety. Even though the evaluation methodology cannot solve the lack of a suitable definition, it can support an approximation of how bit safety is best met.

The need for some sort of an *evaluation methodology* has been addressed in literature, both in questions of bit safety definition, and in questions of how to choose a bit preservation solution, where additional requirements concerning costs and confidentiality are taken into account. Also the case study of the IR-BR pointed to the need for differentiated requirements for a bit preservation solution in accord with the purpose of the digital material for the bit preservation users, including information security and cost requirements.

An *evaluation* of a bit preservation strategy needs to be based on requirements for the bit preservation. Use of the *representation concept* to analyse the digital material for bit preservation gives some of these requirements, since the analysis will reveal requirements for the use and preservation of a representation, which can be requirements for services that need to be offered by the bit repository where the representation is placed. One of the examples was mass processing services for access and migration actions.

Another point from the *representation concept* is that a final decision on a bit preservation representation may not be possible solely on an analysis based on the representation concept. It may for instance need to be assisted by a joint *evaluation* of possible bit preservation representation and possible bit preservation solutions. Thus the representation concept calls for means to do such evaluation as well.

The *evaluation methodology* also intends to include cost requirements. Costs have been addressed as an important aspect in evaluation of choices of a preservation format for a file in a bit preservation representation, or in choice of a bit preservation solution. Evaluation includes economic aspects where a bit preservation strategy respects the budgets set for the entire preservation. Additionally there is the case where costs can be affected by the choice of a digitisation process, and thus early consideration of

representation before a digitisation can be important for the final optimal choice of bit preservation representation and bit preservation solution.

The *evaluation methodology* enables breaking down the unsolved problem of defining bit safety measures and similarly for confidentiality, costs, and availability. The results of evaluation are not precise measures for e.g. requirements for bit safety, but the methodology provides both approximation and documentation of the measures and covers a wide range of interrelated and sometimes conflicting requirements. It is especially the introduction of the BR-ReMS in the evaluation methodology which enables a breakdown of the requirement into requirements expressing different parts of the challenges with bit safety and conflicting requirements.

An important concept in use of the *evaluation methodology* is the SLA, which is an agreement between users and providers of a bit repository. It is the SLA that defines the number of replicas and where they are placed, and aspects like the frequency of integrity checks. Thus the question of how many replicas of data needed is formulated indirectly via the SLA. The evaluation methodology provides means to approximate what a SLA must cover in order to fulfil the various requirements, where the coverage includes the actual pillars used.

The *evaluation methodology* can assist in decision taking for bit preservation of digital material in different stages of the materials life cycle. At an early point, it can be useful as part of analysis of a digitisation process, secondly in choice of a bit preservation solution, and thirdly as part of re-evaluation of the bit preservation strategy. However, as argued, continuous evaluation of sustainable bit repository solutions must be supported by other means of auditing tools. The evaluation methodology can only assist in re-evaluation needed when requirements or conditions for the bit preservation solution have changed. It can never contribute to trust in the actual bit preservation solution or in claimed values of characteristics, on which the evaluation methodology is based.

In the scope of *the holistic approach*, one of the main goals is to optimise the bit preservation strategy respecting various requirements coming from the “whole”. The *evaluation methodology* supports the choice between possible bit preservation strategies expressed in terms of bit preservation solutions. The evaluation does, however, not make sense unless bit preservation can be defined as a separate task. Thus the *IR-BR model* is needed for this purpose. Furthermore, evaluation will need to be based on requirements from the “whole”, which need to be analysed based on requirements from the “whole” to the bit preserved material. Thus the *representation concept* is needed for this purpose. Consequently, the main conclusion is that all three main results are important contributions to finding the optimal bit preservation strategy, and all contributions fill gaps previously left open.

7. Further work

There are various areas in relation to the results of this thesis which call for further work. This includes specific refinement of the results as well as areas not yet dealt with in the holistic approach to bit preservation.

There is still work to be done on the representation concept. The current version is insufficiently detailed and is primarily based on observations of how dissemination and functional preservation requirements affect representation for bit preservation. In this study different issues were found that need further investigation, for example referencing into digital material. There may be other areas that need special attention in representation for digital material with other purposes, such as e-science material. Thus further work needs to be done both with the concrete issues and with broadening the scope of different purposes that can affect requirements for bit preservation, and thus the scope of requirements to be considered in a holistic approach to bit preservation.

Tools and the use of tools in the evaluation methodology also deserve improvements. The individual calculation functions of probability fulfilment need refinement, and some of calculation functions could be the subject for specific new research topics. One way could be to detail the functions based on the current approach by dividing the calculations into smaller problems. Another way could be to consider using Bayesian statistics which can include assumptions of distribution of events.

Specifically, the specifications in the BR-ReMS still lack specification of availability, including mass processing, and costs. These specifications must be incorporated as well. One way to incorporate costs could be to use models like the one described in “Cost Model for Digital Curation: Cost of Digital Migration” [77].

Other elements in the evaluation methodology can also be refined. One aspect is refinement of existing breakdown of overall requirements. For example requirements to cover importance to avoid mass loss could be specified. Another aspect is that use of utility analysis can be improved by more advanced use of weights in use of the Plato tool. Finally, the BR-ReMS specifications can be considered to be extended with other perspectives of a holistic approach, as for example fulfilment green IT requirements.

As mentioned, the evaluation methodology can be used for choosing between representations for bit preservation and different digitisation processes. However, such evaluation would require means to compare outputs from specific evaluations for each representation. Thus, before the evaluation methodology can be used in such cases, more work on how to include such comparisons would be needed. Finally, the role and contents of SLAs are essential for both definition and follow up on bit preservation solutions. Thus means to define such SLAs could be a subject for further studies as well.

There are also aspects directly concerned with how a bit repository can be implemented in a way which can meet different requirements for bit preserved material. One of these aspects is for instance mass processing.

Since my back ground is that of a computer scientist, the computer science aspects are those that have mainly been treated in the holistic approach to bit preservation. There are many other aspects, like economy, which could be much more elaborated, for instance to take into account sustainability aspects of continuous funding which include periods of economic recession. Furthermore, there are several of the functions in the BR-ReMS which could be improved on basis of statistical methods, and knowledge from disciplines such as geology and warfare.

8. Terminology and abbreviations

Terminology is important for common understanding. Especially in a field like digital preservation where e.g. libraries and archives seeks to join forces on an international level to meet the challenges, and where digital preservation covers a wide range of areas where different educational skills are needed. These differences open up for different interpretations of terms due to difference in cultural, language, and educational background. There are reference models and initiatives that have made an effort and played major roles in to definition of common terms, e.g. the OAIS reference model. However, these terms are not always the ones used in practice.

The papers, that this thesis is based on, have been made in different contexts and at different points in time. There are therefore examples of terms that are used differently in the different papers. The most important terms regarding different interpretations are:

- **Manifestations contra representation:** In the first versions of the logical object model in the Planets project, manifestation was used for what is defined as representation in this thesis. However, the term manifestation is also used in IFLA in a more broad interpretation, where e.g. a manifestation of a book can be a theatre play [141]. Paper C is using the terminology from the Planets logical object model, where manifestation is used for a representation. The later Paper A uses instead defines the term representations, and instead uses the term manifestation as it is defined in IFLA. In this thesis manifestation is used in the IFLA interpretation. The thesis uses the definition of representation as defined in Paper A.
- **Meaning of Strategy:** Paper A and Paper B is based on a specific study within the Royal Library of Denmark, and uses the term strategy as an overall strategy of how to achieve a goal, while Paper E which uses Plato, is based on the more concrete definition of strategy of how to achieve a goal with precise solutions with specific tools and implementations. In this thesis the term strategy will either be specified as overall, or it will be prefix with the type of strategy in question e.g. using an emulation strategy. Only in context of the evaluation methodology, the strategies are specifically referring to specific solutions.
- **Migration contra transformation:** The term migration is related to use of a migration strategy of the functional preservation. In Paper A the word transformation is used a few times in connection with changes in representation that are not directly related to migration as part of a migration strategy. This thesis uses the same terminology as defined in Paper A. The only exception is media migration which refers to copying data from one media to another (called refreshment in OAIS terminology). It should be noted that the way migration and transformation is used in this thesis is opposite to OAIS terminology which uses migration in a broader sense [16].
- **Functional preservation contra logical preservation:** These two terms are synonymous. The term functional preservation is the term used in this thesis and in papers related to case studies at the Royal library, i.e. Paper A, Paper B and Paper D. The term logical preservation is used in Planets related research, i.e. Paper C and Paper E.

The following is the list of terms, acronyms and abbreviations used in this thesis:

Access: is the OAIS functional entity that contains the services and functions which make the archival information holdings and related services visible to Consumers. Further description can be found in the Reference Model for an Open Archival Information System (OAIS) [16].

Active Bit preservation: is bit preservation where copies of data are equally worthy copies that are actively checked for integrity and existence on a regular basis.

ADL: is short for Archive of Danish Literature, which is the name of the website which was basis for case studies in Paper A and Paper B.

AIP: is an abbreviation of *Archival Information Package*. Please refer to explanation there.

Administration: is the OAIS entity that contains the services and functions needed to control the operation of the other OAIS functional entities on a day-to-day basis. Further description can be found in the Reference Model for an Open Archival Information System (OAIS) [16].

Analogue material: is a continuous representation of information, e.g. an image recorded on film or sound recorded on a long-playing record. Digital information, on the other hand, represents materials using only ones and zeros.

Analogue preservation: is preservation of analogue material.

Archival Information Package (AIP): is an OAIS term for an information package which is preserved. It includes information in order to make it an individual package and includes preservation information. A more precise description can be found in the Reference Model for an Open Archival Information System (OAIS) [16].

Archival Storage: is the OAIS function entity that contains the services and functions used for the storage and retrieval of Archival Information Packages. Further description can be found in the Reference Model for an Open Archival Information System (OAIS) [16].

ARK: is short for Archival Resource Key, which is the name of a naming scheme for persistent identifiers, where the founding principle of the ARK is that persistence is purely a matter of pointing to a service which can provide the required object [83].

Audit trail: is the information which includes all history of access of the object (e.g. when giving read access to the object), changes in object (e.g. as a consequence of a correction), or changes for the object (e.g. if the media for the object is changed).

Authenticity: is defined as the degree to which a person (or system) may regard an object as what it is purported to be [16,17,45].

Availability: is the property of being accessible and usable upon allowed demand. Further information can be found in the ISO 27000 series [68]

Backing up: is creation of a backup copy of data. See further description under the Backup term.

Backup: denotes a copy of data separately from the original data, so that this additional copy may be used to restore the original after a data loss event on the originals. Further information can also be found in "Bit Preservation: A Solved Problem?"[143].

Baqt: is a file packaging format for storage and transfer of arbitrary digital content. A "bag" has enclosed descriptive tags [84].

Bayesian statistics: is based on work by Thomas Bayes (1702-61). Calculations in Bayesian statistics are using the Bayes theorem based on assumptions or knowledge of data distributions which is formalised as priori distribution. The calculations results in a posterioris distribution of the parameters, based on the priori distribution and information in the data [29].

BBC: is the non-commercial British radio and television broadcasting company.

Bit: is a binary digit that can have value 0 or 1.

Bit integrity: denotes the integrity of bit-streams, which in practice means that the bit-streams are intact.

Bit Integrity check: is a check of whether the bit integrity is intact in all replicas.

Bit preservation: is preservation of bit-streams. *Bit preservation* is defined as the required activities to ensure that the bit-streams remain intact and readable. The paper “Thirteen Ways of Looking at...Digital Preservation” defines it as “... an assurance that the bit streams constituting the digital objects remain intact and recoverable over the long-term.” [86], which basically means the same.

Bit repository (BR): is a repository concerned with long-term preservation of bit-streams, i.e. it is not directly concerned with functional preservation.

Bit Repository – Requirement Measuring System (BR-ReMS): is the name of the tool used for calculation of probabilities for requirement fulfilment by a bit preservation solution. The tool is used in the evaluation methodology described in Paper E and chapter 4.

Bit safety: denotes how well the bits are secured or safeguarded against failures that can damage the bit integrity or the readability.

Bit-stream: is a delimited sequence of bits. All digital material is basically based on bit-streams, for instance a file is basically represented as bits in a bit-stream.

BR: is an abbreviation of *bit repository*. Please refer to explanation there.

BR-ReMS: Abbreviation of *Bit Repository – Requirement Measuring System*. Please refer to explanation there.

Broad crawl: is a harvest, of the World Wide Web, that attempts a broad, automatic sampling of all web pages that are included in the broad call. This is defined as a snapshot harvest in the paper “Overview of the Netarkivet web archiving system” [22].

Characteristic: is description of an aspect as a property/value pair [26]

Characterisation: involves extraction of information that describes the digital object, e.g. metadata, file format features and aspects of content. Further description can be found in the introduction of this thesis and e.g. in the paper “Systematic Characterisation of Objects in Digital Preservation” [7].

Checksum replica: is a replica that only contains the checksum of the data replicated.

Chronopolis: is a system to provide services for the long-term preservation. It is created by The San Diego Supercomputer Center, UC San Diego Libraries with the National Center for Atmospheric Research and the University of Maryland Institute for Advanced Computer Studies. The creation was originally funded by the Library of Congress. Chronopolis includes geographically distributed data and some auditing. Further information can be found on <https://chronopolis.sdsc.edu/>.

Clouding services: are services that can be offered by cloud computing. According to the U.S. National Institute of Standards and Technology cloud computing is defined as being a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources, which can be found on <http://csrc.nist.gov/nice/states/maryland/posters/cloud-computing.pdf>. Clouding services and cloud computing is however not a well-defined concept as discussed in the article “The Internet Industry Is on a Cloud -- Whatever That May Mean” from the Wall Street Journal, March 26, 2009, <http://online.wsj.com/article/SB123802623665542725.html>.

Confidentiality: is the property that information is not made available or disclosed to unauthorised individuals or processes. Further information can be found in the ISO27000 series [68].

Consumer: is an OAIS role, which is played by those persons or client systems, which interact with OAIS services to find preserved information of interest and to access that information in detail. This can include other OAIS systems, as well as internal OAIS persons or systems. Further description can be found in the Reference Model for an Open Archival Information System (OAIS) [16].

Costs: is understood in terms of money.

DAT: is short for Digital Audio Tape is a signal recording and playback medium developed by Sony and introduced in 1987.

Data Management: is an OAIS function entity that contains the services and functions for populating, maintaining, and accessing a wide variety of information. Further information can be found in the Reference Model for an Open Archival Information System (OAIS) [16].

DCC: is an abbreviation of *Digital Curation Centre*. Please refer to explanation there.

Deduplication: refers generally to eliminating duplicate or redundant information. A deduplication system looks for repeating patterns of data at the block and bit levels. When multiple instances of the same pattern are discovered, the system stores a single copy of the pattern [8].

DELOS: was a Network of Excellence on Digital Libraries partially funded by the European Commission. The main objectives of DELOS were research, whose results were in the public domain, and technology transfer, through cooperation agreements with interested parties [28].

Detailed strategy: is a strategy which is very detailed, for example in choosing a specific bit preservation solution including a specific service level agreement.

DIAS: is the IBM Digital Information Archiving System currently used at National Library of the Netherlands and in the KOPAL system. Further information can be found on the DIAS website [30].

Digital Curation Centre (DCC): serves as national centre for solving challenges in digital curation that could not be tackled by any single institution or discipline in the United Kingdom. DCC has e.g. contributed to evolution of DRAMBORA, and hosts the International Digital Curation Conferences. Further information can be found on their website <http://www.dcc.ac.uk/>.

Digital information: is general denotation of information consisting of bit-streams.

Digital material: is material in digital form. It can be synonymous to Digital Object in the definition of as an object composed of a set of bit sequences. The term is used in order to clarify that it is a representation of material, where object can be more loosely understood.

Digital object: is defined as an object composed of a set of bit sequences in OAIS [16]. However, it can have different interpretations in different contexts on exactly what defines a set of bit sequences to be a digital object. In many cases a digital object is synonymous to what is defined as digital material in this thesis.

Digital preservation: is preservation of digital material. It designates the methods and systems which are needed to ensure access to digital materials over time. It covers the series of managed activities necessary to ensure continued access to digital materials for as long as necessary [31,72].

Digital Preservation Coalition (DPC): is primarily a coalition for the United Kingdom, but it also covering internationally. It is a non-profit membership organisation in the United Kingdom whose primary objective is to raise awareness of the importance of the preservation of digital material and the attendant strategic, cultural and technological issues. Further information can be found on <http://www.dpconline.org/>.

Digital substitution: of analogue material is a special case of digitisation, where the purpose of the digitisation is to create a digital copy of a non-digital material in order to support preservation [75].

Digitally born material: is digital material that has been created as digital, i.e. it has no origin in analogue forms. An example of a digitally born material is an email.

Digital Preservation Europe (DPE): was an initiative that fosters collaboration and synergies between many existing national initiatives across the European Research Area. DPE carried out an analysis of all existing research agendas in order to sum up what had to be done and identify missing aspects of the

problem which ultimately led to the foundation of Europe's research and development in terms of digital preservation [167].

Digitisation: is the name of the process where analogue material is transcribed to digital material which can represent the analogue material in a digital form.

Digitised book: is a book that is digitised. Usually a book is digitised by scanning the pages, and possibly extracting OCR and subsequent encoding.

Digitised material: is analogue material transcribed to digital material which can represent the analogue material in a digital form.

DIP: is an abbreviation of *Dissemination Information Package*. Please refer to explanation there.

Dissemination: is the process of providing a DIP to a consumer in OIAS terms. It is here used more broadly in order to avoid explicit reference to OAIS terms like Data Dissemination Session. More information can be found in the Reference Model for an Open Archival Information System (OAIS) [16].

Dissemination Information Package (DIP): is an OAIS term for an information package which is derived from one or more AIPs, received by the Consumer in response to a request to the OAIS. Further information can be found in the Reference Model for an Open Archival Information System (OAIS) [16].

DPC: is an abbreviation of *The Digital Preservation Coalition*. Please refer to explanation there.

DPE: is an abbreviation of *Digital Preservation Europe*. Please refer to explanation there.

DRAMBORA: is a tool for internal audit of digital repositories named: Digital Repository Audit Method Based on Risk Assessment. Further information can be found on <http://www.dcc.ac.uk/resources/tools-and-applications/drambora>, in the DRAMBORA documentation [27] and on the DRAMBORA website [32].

DSpace: is name of open source software for open digital repositories. DSpace based on retrieved metadata for a database based metadata store and production procedures including backup and disaster recovery plans [33]. DSpace has a community around it, and is hosted by the DuraSpace organisation.

DuraCloud: is a hosted service and open technology that enables use of cloud services in a bit preservation solution. DuraCloud is based on open source software, it has a community around it, and is hosted by the DuraSpace organisation. Further information can be found on the DuraCloud website [34].

DuraSpace: is an independent not-for-profit organisation, which focuses on open technologies that provide long-term, durable access to digital assets of scholarly, scientific and cultural data. It covers different open source repository solutions like DSpace, Fedora Commons and DuraCloud. More information can be found on <http://www.duraspace.org/>.

DVD: is short for Digital Versatile Disc or Digital Video Disc. It is a high-density optical disk for storing of data, especially high-resolution audio-visual material.

E-science: is not an unambiguous concept, but can be said to cover research which usually produces very large amounts of electronic data, often performed over distributed networks. Taken from <http://kubis.ku.dk/videncenter/english/about/e-science-data/>.

Emulation: denotes a digital preservation action or strategy, where data is rendered in a new environment via the emulated environment of the old environment. Further information can be found the KEEP website [74], the paper "Ensuring the longevity of digital information" [147], or the paper "Requirements for Applying Emulation as a Preservation Strategy" [181].

Ex Libris: is a commercial company providing library systems solutions [36].

Fedora: is short for *Flexible Extensible Digital Object Repository Architecture*, and is the short name for *FedoraCommons*. Please refer to explanation there.

FedoraCommons: is a community forum and a Repository Project built on open source. The open source software is forming an architecture for storing, managing, and accessing digital content. The Fedora Repository Project and the Fedora Commons community forum are under the stewardship of the DuraSpace organisation [41]. Fedora has been used in a wide range of repositories and systems, for example eSciDoc and Hydra [60,137]. Further information can be found on <http://www.fedora-commons.org/about>.

File format: denotes the representation of digital information in the way that information is encoded for in a computer file. The file format defines how a representation in the format can be interpreted in a presentation of the representation. Examples of file formats are TIFF and GIF.

FineReader: is a tool for OCR (optical character recognition) for text recognition. Further information can be found on <http://finereader.abbyy.com/>.

Full replica: is a replica that contains a full copy of the data replicated. The term is used to spell out that it is not a checksum replica.

Function entity: is an OAIS term, used for entities in the OAIS reference model as illustrated in Figure 6. Further details can be found in the Reference Model for an Open Archival Information System (OAIS) [16].

Functional preservation: is the part of digital preservation that ensures that the bits remain understandable and usable according to preservation purpose.

GIF: is a digital bitmap image format. GIF is an abbreviation of *Graphics Interchange Format*.

Green IT: refers to use of computers and related resources in a manner that includes environmentally issues [49,161].

Handle: offers resolution services for unique and persistent identifiers of digital object [171].

Holistic: originates from holism which is the theory that whole entities, as fundamental components of reality, have an existence other than as the mere sum of their parts [189].

HTML: is the mark-up language for web pages. HTML is an abbreviation of *HyperText Markup Language*. Further information can e.g. be found on <http://tools.ietf.org/html/rfc2854>.

IBM: is a multinational computer, technology and IT consulting corporation.

IFLA: is an abbreviation of *International Federation of Library Associations and Institutions*. Please refer to explanation there.

IIPC: is an abbreviation of *International Internet Preservation Consortium*. Please refer to explanation there.

Ingest: is the OAIS function entity that contains the services and functions that accept SIPs from Producers, prepares AIPs for preservation. More detailed information can be found in the Reference Model for an Open Archival Information System (OAIS) [16].

Institution Repository (IR): is a repository concerned with long-term preservation of digital material in an institution.

Integrity: is the property of safeguarding the accuracy and completeness of digital material. Further information can be found in the ISO27000 series [68].

International Federation of Library Associations and Institutions (IFLA): is an international body representing the interests of library and information services and their users. Further information can be found on <http://www.ifla.org/en/about>.

International Internet Preservation Consortium (IIPC): is a Consortium with focus on internet preservation. The mission of IIPC is to acquire, preserve and make accessible knowledge and information from the Internet for future generations, promoting global exchange and international relations. The goals are to enable the collection, and support in Internet archiving and preservation partly by fostering development and use of common tools [63].

IR: is an abbreviation of *institution repository*. Please refer to explanation there.

IR-BR model: is a model that separates bit preservation in a bit repository (BR) as part of an institution repository (IR). The model is described in Paper D and chapter 2.

iRODS: is an acronym of the *Integrated Rule Oriented Data System*, which is an open source data grid software system originally developed by the Data Intensive Cyber Environments research group, and collaborators. It is a rule oriented software system, which enables specification of management policies, the Rule Engine of iRODS interprets the Rules to decide how the system is to respond to various requests and conditions. Further information can e.g. be found on the iRODS website [62] or in the paper “Management and Preservation of Research Data with iRODS” [52].

ISO: is short for *International Organisation for Standardization*, which develops and develops and publishes international Standards. It is a non-governmental organisation with a network of the national standards institutes of a central secretariat that coordinates the system. Further information can be found on <http://www.iso.org/iso/about.htm>.

JHOVE: is a validation tool which provides functions to perform format-specific identification, validation, and characterisation of digital objects. Further information can be found on <http://hul.harvard.edu/jhove/> and in the paper “‘What? So What’: The Next-Generation JHOVE2 Architecture for Format-Aware Characterization” [1].

JPEG: is a digital image format with lossy compression. JPEG is an acronym for the *Joint Photographic Experts Group* which created the standard.

JPEG2000: is a digital image format which can be with lossy or lossless compression. JPEG is an acronym for the Joint Photographic Experts Group which created the standard in year 2000. Further information can e.g. be found on <http://tools.ietf.org/html/rfc3745>.

KEEP: is an acronym for *Keeping Emulation Environments Portable*, which is an EU project developing emulation services to enable rendering of both static and dynamic digital objects. The overall aim of the project is to facilitate universal access to our cultural heritage by developing flexible tools for accessing and storing a wide range of digital objects [167].

KOPAL: is a German archival system. Further information can be found on the KOPAL website [79].

LOCKSS: is an acronym for *Lots of Copies Keep Stuff Safe*, which is an international community initiative that provides libraries with digital preservation tools and support. LOCKSS provides a peer-to-peer, decentralised digital preservation infrastructure open source tool which focus on easy and inexpensive collection and preservation of e-content. Further information can e.g. be found on <http://www.lockss.org/lockss/Home>, or in the papers “The LOCKSS Peer-to-Peer Digital Preservation System” [96] or “LOCKSS (Lots of copies Keep stuff safe)”.

LOCKSS cache: is an independent unit in LOCKSS corresponding to a pillar in the general view of a bit repository illustrated in Figure 5. Further information can be found in the documentation of LOCKSS.

Logical preservation: is synonymous to *functional preservation*. See explanation there.

LZW compressed: denotes that a file is compressed based on the LZW lossless data compression algorithm. LZW is short for *Lempel–Ziv–Welch* which are the names of the creators of the algorithm. Further information can be found on:

http://www.cs.duke.edu/courses/spring03/cps296.5/papers/welch_1984_technique_for.pdf

Management: is the role played by those who set overall OAIS policy as one component in a broader policy domain. Further information can be found in the Reference Model for an Open Archival Information System (OAIS) [16].

Manifestation: has more meanings depending on the context. In the first versions of the logical object model in the Planets project, manifestation was used for what is defined as representation in this thesis. This meaning of the concept is only used in Paper C. The term manifestation is also used in IFLA in a more broad interpretation, where e.g. a manifestation of a book can be a theatre play [141].

Many-to-many relation: is a specific kind of relation between two objects. In case the objects are tables in a database, it means that one record in either table can relate to many records in the other table.

Mass digitisation: covers large-scale projects that digitise analogue material. In recent years there have been many examples of mass digitisation of books. An example can be found for digitisation of Danish national literature from before year 1700 [119]. A definition of mass digitisation can be found in the paper “The Seamless Cyberinfrastructure: The Challenges of Studying Users of Mass Digitization and Institutional Repositories” [155].

Mass processing: means processing of large data archives, as e.g. a bit repository. The term is analogously to mass digitisation. An example of mass processing can be found for the Danish web archive of 71 TB of data in 2008 in order to find mime types of file in the archive. A mass processing on this archive is the basis for the statistics presented on page 13 of the paper “Erfaringer med høstning af det danske net 2005 – 2008” [70].

MD5 checksum: is a checksum calculated on basis of the MD5, where MD5 is short for Message-Digest algorithm 5. Further information can be found on <https://tools.ietf.org/html/rfc1321>.

Media migration: denotes migration of data from one media to another by copying the data. Media migration is also sometimes referred as hardware migration, or in OAIS terms as refreshment.

MetaArchive: is a digital preservation network created and hosted by and for cultural memory organisations. It is a community-owned, community-led initiative comprised of libraries, archives, and other digital memory organisations. Working cooperatively with the Library of Congress through the NDIIPP Program, aiming at a secure and cost-effective repository that provides for the long-term care of digital materials during actively participating in the preservation of institutions own content [100].

Metadata: is loosely defined as data about data. There are different types of metadata, for instance preservation metadata as in PREMIS [132] and descriptive, administrative, and structural metadata as in METS [101].

METS: is short for *Metadata Encoding & Transmission Standard*, which is a standard for encoding descriptive, administrative, and structural metadata regarding objects within a digital library [102]. METS provides XML schemas for the standard. Further information can be found on the METS website [101] and in the standard [102].

Micro service: is a small procedure that perform a certain task and which is made available for the iRODS server code [103]. Micro services also relates to the micro service concept described in “An Emergent Micro-Services Approach to Digital Curation Infrastructure” [163]

Migration: Except from media migration, migration denotes a digital preservation action or strategy, where data is migrated in order to be accessible with new technology. The corresponding transformation action defined in the OAIS reference model is defined as the transfer of digital information, while intending to preserve it [16]. In other words migration consists of migrating data from one representation to another, i.e. from one structure represented in a set of files to a possibly new structure and a new set of files with new formats. Further information can be found in the report “Long-term Preservation of Web Archives – Experimenting with Emulation and Migration Methodologies” [92], the paper “Ensuring the longevity of digital information” [147], or in the paper on “The Planets Approach to Migration Tools” [194].

MIX: is short for *Metadata for Images in XML*, which is a standard for metadata to still images. MIX provides XML schemas for the standard [108].

MODS: is short for *Metadata Object Description Schema*, which is a standard for encoding descriptive metadata. MODS provides XML schemas for the standard [109].

National Digital Information Infrastructure & Preservation Program (NDIIPP): is a digital preservation network that connects libraries, archives, universities, research centres, non-profit and for-profit organisations and professional associations across the United States and the rest of the world. The focus for the program was; capturing, preserving, availability of digital content, building networks and developing a technical infrastructure. Further information can be found on their website <http://www.digitalpreservation.gov/>.

National Digital Stewardship Alliance (NDSA): is initiated by NDIIPP, and is a member based alliance aimed for organisations that are interested in enduring access to digital information. The alliance gives members access to expert information about current practices, tools and services. The NDSA alliance is an initiative to reach out to all institutions with interest and obligation to do digital preservation [112].

NCDD: is an abbreviation of *The Netherlands Coalition for Digital Preservation*. Please refer to explanation there.

NDIIPP: is an abbreviation of *National Digital Information Infrastructure & Preservation Program*. Please refer to explanation there.

NDSA: is an abbreviation of *National Digital Stewardship Alliance*. Please refer to explanation there.

nestor: is short name for *The German Network of Expertise in long-term storage of digital resources*. Please refer to explanation there [115].

Netarchive.dk: is the name of the Danish web archive, which collects and preserves the Danish portion of the internet. Further information can be found on the Netarchive.dk website [116]. In 2005 there were made a very rough estimate of the how high a pile of paper would be if one Terabyte of harvested data was printed, which gave 20 a km pile. This number is subject to high uncertainty, and it only has an internal reference, apart from verbal presentations, e.g. in an interview on the Danish national television in 2009.

Netherlands Coalition for Digital Preservation (NCDD): is a coalition that covers the public sector of the Netherlands. It aims for a sustainable technical and organisational infrastructure. Furthermore it acts as a catalyst and joint platform for sharing expertise and advocacy issues. Further information can be found the website of the coalition [118].

NFS: is an abbreviation of the *United States National Science Foundation*. Please refer to explanation there.

Normalised form: is in this thesis used by the meaning of a standardised form that minimises loss of information and ensures consistent use and maintenance of similar structures.

OAIS: is an abbreviation of Open Archival Information System, see explanation there.

OCLC: is an abbreviation of the *Online Computer Library Center*. Please refer to explanation there.

OCR: is an abbreviation for Optical Character Recognition. It is digital translation of scanned images of printed text into encoded text.

Online Computer Library Center (OCLC): is a non-profit, membership, computer library service and research organisation dedicated to the public purposes of furthering access to the world's information and reducing the rate of rise of library costs. The RLG was merged with the OCLC in 2006 [51]. Further information can be found on <http://www.oclc.org/uk/en/about/default.htm>.

Opaque identifier: is an identifier with opaque contents, i.e. it is not possible to reveal any humanly understandable information from the identifier, e.g. in case it is an identifier for library information, the identifier will *not* contain information about the library etc. Opaque identifiers are for instance discussed in "*Persistent Identifiers: Considering the Options*" [178].

Open Archival Information System (OAIS): is an archive, consisting of an organisation of people and systems that has accepted the responsibility to preserve information and make it available for a user. More detailed information can be found in the Reference Model for an Open Archival Information System (OAIS) [16].

Open Planets Foundation (OPF): is a member based foundation, which has been established to provide practical solutions and expertise in digital preservation, building on the research and development outputs of the Planets project [124]. Further information can be found on the OPF website [124].

OPF: is an abbreviation of *Open Planets Foundation*. Please refer to explanation there.

Overall strategy: is a strategy which is not detailed in what tools and solutions to choose. An example of an overall strategy is an emulation strategy for web materials.

PADI: is an abbreviation of *Preserving Access to Digital Information initiative*. Please refer to explanation there.

PDF/A: is a standardised electronic document file format for long term preservation [65,67].

Peer-to-peer: is used for systems that consist of a number of peer systems, where there typically is lacks of a dedicated, centralised infrastructure, but depend on the voluntary participation of peers to contribute resources out of which the infrastructure is constructed [153].

Persistent identifier: is not a well-defined concept, but is usually used for identifiers that cannot be reused for other digital objects than the object that it was originally assigned to. The ambiguity relates to definition of the object or objects that the identifier is assigned to as well as the placement of them, the difficulty being social rather than technological [44, 178].

Pillar: is the unit that forms the basic storage for a copy of data in a bit repository. A pillar is a unit based on specific technology with an organisation around it which e.g. can be responsible for operation, technology watch etc.

Planets: is an acronym for *Preservation and Long-term Access through Networked Services*, see explanation there.

Plato: is the name of a preservation planning tool used in the Planets project and in the evaluation methodology described in Paper E and chapter 4. Further description can be found in the paper "*Plato: A service oriented decision support system for preservation planning*" [6] and in Appendix I "*Detailed Calculations using Plato*".

PLN: is an abbreviation of *Private LOCKSS Networks*. See explanation there.

PREMIS: is an acronym for Preservation Metadata: Implementation Strategies, which is a Data Dictionary for preservation metadata [132]. PREMIS provides XML schemas for the standard. Further information can be found on the PREMIS website [131] and in the PREMIS documentation [132].

Preservation action: is an action taken as part of preservation. In this thesis this term is used in connection with description of functional preservation in the perspective of the Planets project. Further information can be found in the paper “Planets: Integrated Services for Digital Preservation” [39], or on the Planets website [128].

Preservation and Long-term Access through Networked Services (Planets): is the name of an EU funded research project on functional preservation running from 2004-2010. The primary goal for the project was to build practical services and tools to help ensure long-term access to digital cultural and scientific assets, with a special focus on the needs of libraries and archives [167]. More information can be found in the paper “Planets: Integrated Services for Digital Preservation” [39], or on the Planets website [128].

Preservation format: is a file format which is chosen to be used in preservation representations.

Preservation level: is the level of preservation for specific digital material. There are no exact definition, but examples can be found in the PREMIS documentation [132] and in the paper “Building Blocks for the new KB E-Depot” [183].

Preservation Planning: is an OAIS function entity that contains the services and functions for monitoring the environment of the OAIS and providing recommendations to ensure that the information stored in the OAIS remains accessible over the long term, even if the original computing environment becomes obsolete. More detailed description can be found in the Reference Model for an Open Archival Information System (OAIS) [16].

Preservation strategy: is the strategy used for preservation. A strategy for functional preservation can e.g. be a migration strategy or emulation strategy, which is an example of an overall strategy. For bit preservation, a strategy can be chosen in form of a specific solution for how the bits are preserved, which is an example of a detailed strategy.

Preservation system: is a system that supports preservation activities.

Preserving Access to Digital Information initiative (PADI): is an organisation originally placed at The National Library of Australia. It aims to provide mechanisms that will help to ensure that information in digital form is managed with appropriate consideration for preservation and future access. The PADI website is a subject gateway to digital preservation resources, with facilitation of exchange of news and ideas about digital preservation issues [133].

Preserving Virtual World: is a project partly sponsored by NDIIPP. The project is a collaborative research venture of a range of US partners. The primary goals the project have been to investigate issues surrounding the preservation of video games and interactive fiction through a series of case studies of games and literature from various periods in computing history, and to develop basic standards for metadata and content representation of these digital artifacts for long-term archival storage [99].

Private LOCKSS Networks (PLN): are networks of specific LOCKSS caches defined by a small private community sharing a LOCKSS network [134,138,139].

Producer: is an OAIS role, which is played by those persons or client systems, which provide the information to be preserved. Further information can be found in the Reference Model for an Open Archival Information System (OAIS) [16].

PRONOM: is an on-line information system about data file formats and their supporting software products. Further information can be found on <http://www.nationalarchives.gov.uk/PRONOM/Default.aspx>.

RAID: is a technology that provides increased storage reliability through redundancy. RAID is an abbreviation of Redundant Array of Inexpensive Disks. Further information can also be found in the paper “A case for redundant arrays of inexpensive disks (RAID)” [126], or in the paper “RAID: High-Performance, Reliable Secondary Storage” [19].

Replica: is a copy of the data stored on a pillar, and that can be seen and analysed as an individual unit at the abstract level.

Repository: is for analogue materials defined as a building or room designed or arranged and used specifically and exclusively for long-term storage of archive and library materials [64]. For digital material this is, however, not bound to building or room. It is instead the organisation and techniques designed and arranged and used specifically and exclusively for long-term storage of digital materials.

Representation: relates to digital material. A representation for digital material consists of files of different formats and structures between the files or fractions of the files. An example is a book which can be represented by images of book pages along with digital information describing the order of the pages. Here the full representation includes all elements that enable the meaningful presentation of the digital material which build on the OAIIS definition of representation information. Further information can be found in Paper A and chapter 3.

Representation Information: is in OAIIS defined as the information that maps a composed set of bit sequences into more meaningful concepts [16].

Requirements tree: denotes requirements described in a mind map tree, which is used as input to the Plato tool used for evaluation of preservation strategies. An example is given in Figure 28. Further description and references can be found in Appendix I “Detailed Calculations using Plato”.

Research Libraries Group (RLG): is an organisation which was merged with the OCLC in 2006 [51]. RLG was founded in 1974 as a not-for-profit membership corporation of universities, libraries, archives, and other institutions supporting research and learning [35].

Risk based approach: refers to an approach of bit preservation where the risks of losing bits are considered. Such a risk based approach can be e.g. be found in the paper “The Requirements for Digital Preservation Systems, A Bottom-Up Approach” presents a detailed list of threats and strategies to avoid threats which again focus on replication, media migration, independence between replicas, and also points at the importance of audits and economy [144]. Further description can be found in section 1.5.1 ‘Bit preservation’.

RLG: is an abbreviation of *Research Libraries Group*, see explanation there.

SCALable Preservation Environments (SCAPE): is the name of an EU funded research project which will focus on following contributions to digital preservation: by developing infrastructure and tools for scalable preservation actions; by providing a framework for automated, quality-assured preservation workflows and by integrating these components with a policy-based preservation planning and watch system [167]. Further information can be found on <http://www.scape-project.eu/>.

SCAPE: is an acronym of *SCALable Preservation Environments*, see explanation there.

SC1: is an abbreviation for the first subcontractor who did re-digitisation as input to Paper B. This subcontractor was the National library of Norway, who uses JPEG2000 with lossless compression as their preservation format. They normally delete the TIFF version of the scanning after processing JPEG2000 and the OCR & encoded texts. Their background for choosing JPEG2000 with lossless

compression is described in the document “Digitization of books in the National Library - methodology and lessons learned” [113].

Service Level Agreement (SLA): is an agreement possibly in form of a contract where the level of service is formally defined. A service level agreement is a negotiated agreement between parties where one is the customer and the others are the service providers [193]. In this thesis the specific definition related to the IR-BR model is: an agreement of level of service between the unit(s) responsible for the BR (the BR provider(s)) and a user preserving bits in the BR (a BR user). There is more discussion about SLAs in section 4.4.5 “Service level agreements”.

Significant properties: are those aspects of the digital material which must be preserved over time in order for the digital object to remain accessible and meaningful. Further information can e.g. be found on the home page for the project information of the Investigating Significant Properties of Electronic Content (InSPECT) project which investigates aspects of significant properties <http://www.significantproperties.org.uk/>.

SIP: is an abbreviation of *Submission Information Package*, see explanation there.

SLA: is an abbreviation of *Service Level Agreement*, see explanation there.

Storage technology: denotes technology concerned with digital storage.

Strategy: refers to a plan of action designed to achieve a particular goal. As described in the introduction of this chapter, a strategy can either be overall or detailed. Please refer to the terms overall strategy and detailed strategy for more information.

Submission Information Package (SIP): is an OAIS term for an information package that is delivered by the producer to the OAIS for use in the construction of one or more AIPs. Further information can be found in the Reference Model for an Open Archival Information System (OAIS) [16].

Substitution copy: is a digital copy of a non-digital material in order to support preservation [75]. In this thesis the term is used in cases where the substitution copy substitutes the original as the master for further preservation.

TAR: is abbreviation of *Tape ARchive*, which are both a packaging file format and the name of a program handling such files e.g. packaging and unpacking the files.

Technological evolution: denotes the evolution of technologies used in connection with digital materials.

Technological museum: is synonymous to *Technology preservation*. See explanation there.

Technology preservation: is a functional preservation strategy which preserves digital material by preserving all needed technical devices, operating system etc. to access the digital material. A more detailed description is given in the report “Long-term Preservation of Web Archives – Experimenting with Emulation and Migration Methodologies” [92]. An early and more comprehensive description of functional preservation strategies can be found in the paper “Ensuring the longevity of digital information” [147].

Technology watch: is the same as the OAIS function *Monitor Technology* under the OAIS functional entity Preservation Planning. This function is responsible for tracking emerging digital technologies, information standards and computing platforms (i.e. hardware and software) to identify technologies which could cause obsolescence in the archive's computing environment and prevent access to some of the archives current holdings. Further description can be found in the Reference Model for an Open Archival Information System (OAIS) [16].

TEI: is an abbreviation of the *Text Encoding Initiative*, which is a consortium that collectively develops and maintains a standard for the representation of texts in digital form. Its chief deliverable is a set of

guidelines which specify encoding methods for machine-readable texts based on the XML format. Further description can be found on the TEI website [169].

TEI-P4: is a specific set of guideline named P4 based on TEI. See further explanation under term TEI and in the Guide lines for P4 [170].

The Digital Preservation Coalition (DPC): is a coalition placed in United Kingdom and is primarily a coalition for the United Kingdom, but it also covering internationally. DPC is a non-profit membership organisation whose primary objective is to raise awareness of the importance of the preservation of digital material and the attendant strategic, cultural and technological issues [15].

The German Network of Expertise in long-term storage of digital resources (nestor): is the German competence network for digital preservation. nestor is a cooperation association including partners from different fields, but all connected to work with digital preservation e.g. libraries, archives, museums and leading experts. More information can be found on <http://www.langzeitarchivierung.de/eng/index.htm>.

TIFF: is a digital image format. TIFF is an abbreviation of Tagged Image File Format. Further information can e.g. be found on <http://tools.ietf.org/html/rfc3302>.

TRAC: is an acronym for Trustworthy Repositories Audit & Certification: Criteria and Checklist, which is an audit and certification tool for trusted digital repositories. More information can be found in the TRAC documentation [122], and on <http://www.dcc.ac.uk/resources/tools-and-applications/trustworthy-repositories>.

Transformation: is not used in the thesis itself, but is used in the papers on which the thesis is based. In Paper A, transformation is used a few times in connection with changes in representation that are not directly related to migration as part of a migration strategy. In Paper C transformation is used in the same meaning as migration. Further description can be found in the introduction to this terminology chapter.

Trust: is used in connection with a trusted digital repository or trustworthy digital repository. According to TRAC a trusted repository is one whose mission is to provide reliable, long-term access to managed digital resources, now and in the future [142].

United States National Science Foundation's (NSF): is an independent federal agency created by Congress in 1950 to promote the progress of science [114].

URI: is an abbreviation for a *Uniform Resource Identifier*, which is a compact string of characters for identifying an abstract or physical representation of a digital object [10].

URN: is an abbreviation for a Uniform Resource Identifier Name. The URNs are subset of URIs and URNs are used for identification, although an URN does not imply availability of the identified digital object [162].

UUID: is a universally unique identifier, which is specified in the ISO/IEC 9834-8 standard. It enables users to produce identifiers that are either guaranteed to be globally unique, or are globally unique with a high probability [69].

WARC: is a standardised storage format, which is aim at web archived material [66].

Web archive: is an archive for collected, archived, and preserved data from the World Wide Web.

XCL: is the name of the *Extensible Characterisation Language*, which is used for comparison of characteristics in digital material before and after a migration [7,172].

XML: is the name of the *Extensible Markup Language*, which is a set of rules for encoding documents in machine-readable form. XML is for instance used for specification of metadata based on metadata schemes such as for MIX, METS and PREMIS.

ZIP: is a file format, which is used for data compression and as an archive format. A ZIP file contains one or more files that have been compressed to reduce file size.

9. References

- [1] Abrams, S., Morrissey, S., Cramer, T.: *“What? So What”: The Next-Generation JHOVE2 Architecture for Format-Aware Characterization*, In: The International Journal of Digital Curation, vol. 4, no. 3 (2009)
- [2] Altenhöner, R.: *E-infrastructure and Digital Preservation: Challenges and Outlook*, In: Proceedings of the 6th International Conference on Preservation of Digital Objects, San Francisco, USA, pp. 12-19 (2009)
- [3] Androutsellis-Theotokis, S., Spinellis, D.: *A survey of peer-to-peer content distribution technologies*, In: Journal ACM Computing Surveys, vol. 36 issue 4 (2004)
- [4] Baker, M., Keeton, K., Martin, S.: *Why Traditional Storage Systems Don’t Help Us Save Stuff Forever*, In: Proceedings of the 1st IEEE workshop on hot topics in system dependability, Japan (2005)
- [5] Baker, M., Shah, M., Rosenthal, D. S. H., Roussopoulos, M., Maniatis, P., Giuli, T. J., Bungale, P.: *A Fresh Look at the Reliability of Long-term Digital Storage*, In: Proceedings of the 1st ACM SIGOPS/EuroSys European Conference on Computer Systems, New York, USA (2006)
- [6] Becker, C., Kulovits H., Rauber A., Hofman H.: *Plato: A service oriented decision support system for preservation planning*, In: Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries, Pittsburgh, USA (2008)
- [7] Becker, C., Rauber, A., Heydegger, V., Schnasse, J., Thaller, M.: *Systematic Characterisation of Objects in Digital Preservation: The eXtensible Characterisation Languages*, In: Journal of Universal Computer Science, vol. 14, no. 18, pp. 2936-2952 (2008)
- [8] Behtash, B.: *Expanding Role Of Data Deduplication*, In: InformationWeek, May 15th 2010, <http://www.informationweek.com/news/storage/systems/showArticle.jhtml?articleID=224701816>, retrieved August 2011 (2010)
- [9] Berman, F., Lavoie, B., Ayris, P., Choudhury, G. S., Cohen, E., Courant, P., et al.: *Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information*, Final Report of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access, http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf retrieved August 2011 (2010)
- [10] Berners-Lee, T., Fielding, R., Irvine, U. C., Masinter, L.: *Uniform Resource Identifiers (URI): Generic Syntax*, <http://www.ietf.org/rfc/rfc2396.txt>, retrieved August 2011
- [11] Beruti, V., Forcada, E., Albani, M., Conway, E., Giaretta, D.: *ESA Plans – A Pathfinder for Long Term Data Preservation*, In: Proceedings of the 7th International Conference on Preservation of Digital Objects, Vienna, Austria, pp. 53-59 (2010)
- [12] Billenness, C. S. G.: *Report on the Proceedings of the Workshop: The Future of the Past – Shaping new visions for EU-research in digital preservation*, http://cordis.europa.eu/fp7/ict/telearn-digicult/future-of-the-past_en.pdf, retrieved August 2011 (2011)
- [13] Boyko, A., Hamidzadeh, B., Littman, J.: *A Framework for Object Preservation in Digital Repositories*, Imaging Science & Technology Reporter, vol. 21, no. 3 (2006)
- [14] Brown D.: *Lost in Cyberspace: The BBC Domesday Project and the Challenge of Digital Preservation*, <http://www.csa.com/discoveryguides/cyber/overview.php>, retrieved August 2011 (2003)
- [15] California Digital Library, <http://www.cdlib.org/>, retrieved August 2011
- [16] CCSDS (Consultative Committee for Space Data Systems): *Reference Model for an Open Archival Information System (OAIS)*, CCSDS 650.0-B-1 Blue book (also published as ISO 14721:2003 in

- updated version), <http://public.ccsds.org/publications/archive/650x0b1.pdf>, retrieved August 2011 (2002)
- [17] CCSDS (Consultative Committee for Space Data Systems): *Reference Model for an Open Archival Information System (OAIS)*, DRAFT RECOMMENDED STANDARD, CCSDS 650.0-P-1.1 Pink Book, <http://public.ccsds.org/sites/cwe/rids/Lists/CCSDS%206500P11/Attachments/650x0p11.pdf>, retrieved August 2011 (2009)
- [18] CCSDS (Consultative Committee for Space Data Systems): *XML Formatted data Unit (XFDU) structure and Construction Rules*, CCSDS 661.0-B-1, Blue book, <http://public.ccsds.org/publications/archive/661x0b1.pdf>, retrieved August 2011 (2008)
- [19] Chen, P. M., Lee, E. K., Gibson, G. A., Katz R. H., Patterson, D. A.: *RAID: High-Performance, Reliable Secondary Storage*, In: ACM Computing Surveys, vol. 26 , issue 2, pp. 145-185, (1994)
- [20] Christensen, N. H.: *A formal analysis of recovery in a preservational data grid*, presented at: The 4th NASA Goddard, 23rd IEEE, Conference on Mass Storage Systems and Technologies, <http://netarchive.dk/publikationer/nhc-kb-dk-msst2006.pdf>, retrieved August 2011 (2003)
- [21] Christensen, N.H.: *Preserving the Bits of the Danish Internet*, In: 5th International Web Archiving Workshop, <http://iwaw.europarchive.org/05/papers/iwaw05-christensen.pdf>, retrieved August 2011 (2005)
- [22] Clausen, L.: *Overview of the Netarkivet web archiving system*, In: Proceeding of the 6th International Web Archiving Workshop, Alicante, Spain, pp. 11-24, Masanès, J., Rauber, A. (eds), <http://www.iwaw.net/06/>, retrieved August 2011 (2006)
- [23] Crespo, A.: *Archival Repositories for Digital Libraries*, Doctoral Dissertation, Stanford University, <http://www-db.stanford.edu/~crespo/publications/thesis.pdf>, retrieved August 2011 (2003)
- [24] Cundiff, M. W.: *An introduction to the Metadata Encoding and Transmission Standard*, In: Library High Tec, vol. 22, no. 1, pp. 52-64 (2004)
- [25] Dappert, A., Enders, M.: *Using METS, PREMIS and MODS for Archiving eJournals*, D-Lib Magazine, vol. 14, no. 9/10 (2008)
- [26] Dappert, A., Farquhar, A.: *Significance Is in the Eye of the Stakeholder*, In: Proceedings of the 13th European Conference on Research and Advanced Technology for Digital Libraries, pp. 297-308, Agosti, M., Borbinha, J. , Kapidakis, S., Papatheodorou, C., Tsakonas, G. (eds.) LNCS, vol. 5714, Springer, Heidelberg (2009)
- [27] DCC (Digital Curation Centre) & DPE (DigitalPreservationEurope), *Digital Repository Audit Method Based on Risk Assessment*, Version 1.0, via <http://www.repositoryaudit.eu/>, requested and retrieved August 2011 (2007)
- [28] DELOS Network of Excellence on Digital Libraries, <http://www.delos.info/>, retrieved August 2011
- [29] Den store danske, Gyldendals åbne encyklopædi (term: *bayesiansk statistik*), Gyldendal, via <http://www.denstoredanske.dk/>, retrieved August 2011 (2011)
- [30] *Digital Information Archiving System*, http://www-935.ibm.com/services/nl/dias/is/implementation_services.html, retrieved August 2011
- [31] *DPC Introduction - Definitions and Concepts*, <http://www.dpconline.org/advice/preservationhandbook/introduction/definitions-and-concepts>, retrieved August 2011
- [32] DRAMBORA interactive, <http://www.repositoryaudit.eu>, retrieved August 2011
- [33] DSpace, <http://www.dspace.org> , retrieved August 2011

- [34] DuraCloud, <http://www.duracloud.org/>, retrieved August 2011
- [35] Erway, E. L.: *Digital Initiatives of the Research Libraries Group*, D-Lib Magazine, December 1996 (1996)
- [36] ExLibris, <http://www.exlibrisgroup.com/>, retrieved August 2011
- [37] Factor, M., Naor, D., Rabinovici-Cohen, S., Ramati, L., Reshef, P., Satran, J.: *The need for preservation aware storage: a position paper*, In: ACM SIGOPS Operating Systems Review archive, vol. 41, issue 1 (2007)
- [38] Factor, M., Naor, D., Rabinovici-Cohen, S., Ramati, L., Reshef, P., Satran, J., Giaretta, D.L.: *Preservation DataStores: Architecture for Preservation Aware Storage*, In: Proceedings of the 24th IEEE Conference on Mass Storage Systems and Technologies, San Diego, California, United States of America, pp. 3-15 (2007)
- [39] Farquhar, A., & Hockx-Yu, H.: *Planets: Integrated Services for Digital Preservation*, In: The International Journal of Digital Curation, vol. 2, no. 2 (2007)
- [40] Fast LTA press release on Silent Cubes, <http://www.fast-lta.de/en/presse/fast-lta-and-medidok/1123>, retrieved August 2011
- [41] Fedora commons, <http://www.fedora-commons.org/>, retrieved August 2011
- [42] *Formal statement of Conformance to ISO 14721:2003*, <http://lockss.stanford.edu/lockss/OAIS>, retrieved August 2011 (2004)
- [43] Gartner, R.: *Metadata for digital libraries: state of the art and future directions*, JISC: Bristol, UK, http://www.jisc.ac.uk/media/documents/techwatch/tsw_0801pdf.pdf, retrieved August 2011 (2008)
- [44] Giaretta, D.: *Advanced Digital Preservation*, Springer, Heidelberg (2011)
- [45] Giaretta, D., Matthews, B., Bicarregui, J., Lambert, S., Guercio, M., Michetti, G. et al.: *Significant Properties, Authenticity, Provenance, Representation Information and OAIS Information*, In: Proceedings of the 6th International Conference on Preservation of Digital Objects, San Francisco, USA, pp. 67-73 (2009)
- [46] Gillesse, R., Rog, J., Verheusen, A.: *Life Beyond Uncompressed TIFF: Alternative File Formats for Storage of Master Image Files*, In: Proceedings of the IS&T Archiving Conference, Bern, Switzerland, pp. 41-46 (2008)
- [47] Gladney, H. M.: *Preserving Digital Information*, Springer, Heidelberg (2007)
- [48] Graham, P. S.: *Intellectual Preservation: Electronic Preservation of the Third Kind*, Commission on Preservation and Access, Washington, D.C. (1994)
- [49] Grindley, N., Negulescu, K. C., Kilbride, W., Macdonald, D., Rosenthal, D.: *Panel discussion: How green is digital preservation?*, In: Proceedings of the 7th International Conference on Preservation of Digital Objects, Vienna, Austria, pp. 217-218 (2010)
- [50] *Gyldendals leksikon*, Gyldendalske boghandel, Nordisk forlag A/S, Copenhagen (1996)
- [51] Hane, P. J.: *RLG to Merge with OCLC*, Information Today, Inc., Posted May 8 2006, <http://newsbreaks.infotoday.com/nbreader.asp?ArticleID=15851>, retrieved August 2011 (2006)
- [52] Hedges, M., Hasan, A., Blanke, T.: *Management and Preservation of Research Data with iRODS*, In: Proceedings of the ACM first workshop on CyberInfrastructure: information management in eScience, Lisbon, Portugal , pp. 17-22 (2007)

- [53] Hedstrom, M.: *Digital Preservation: A Time Bomb for Digital Libraries*, In: Computers and the Humanities, vol. 31, no. 3, pp. 189-202, Springer, Heidelberg (1997)
- [54] Heydegger, V.: *Just One Bit in a Million: On the Effects of Data Corruption in Files*, In: Proceedings of the 13th European Conference on Research and Advanced Technology for Digital Libraries, pp. 315-326, Agosti, M., Borbinha, J., Kapidakis, S., Papatheodorou, C., Tsakonas, G. (eds.) LNCS, vol. 5714, Springer, Heidelberg (2009)
- [55] Hodge, G.: *Preservation of and permanent access to electronic information resources: A system perspective*, In: Journal of Information Services and Use, vol. 25, no. 1, pp. 47-57 (2005)
- [56] Hofmann, A., Giel, D. M.: *DANOK: Long Term Migration Free Storage of Digital Audio Data on Microfilm*, In: Proceedings of the IS&T Archiving Conference, Bern, Switzerland, pp. 184-187 (2008)
- [57] Holley, R.: *Many Hands Make Light Work: Public Collaborative OCR Text Correction in Australian Historic Newspapers*, Technical Report from National Library of Australia, http://www.nla.gov.au/ndp/project_details/documents/ANDP_ManyHands.pdf, retrieved August 2011 (2009)
- [58] *Hoppla 2.0 User Guide Version 0.5*, http://www.ifs.tuwien.ac.at/dp/hoppla/release/download/Hoppla_User_Guide_2.0.pdf, retrieved August 2011
- [59] Hyatt, S., Connaway, L. S.: *Utilizing E-books to Enhance Digital Library Offerings*, In: Ariadne Web Magazine, issue 33 (2002)
- [60] Hydra, <https://wiki.duraspace.org/display/hydra/The+Hydra+Project>, retrieved August 2011
- [61] IBM Cloud Computing, http://www.ibm.com/ibm/cloud/?cm_re=masthead- -solutions- -cloud, retrieved August 2011
- [62] Integrated Rule-Oriented Data System, <https://www.irods.org>, retrieved August 2011
- [63] International Internet Preservation Consortium, <http://netpreserve.org/about/index.php>, retrieved August 2011
- [64] ISO 11799:2003, *Information and documentation -- Document storage requirements for archive and library materials*, via http://www.iso.org/iso/iso_catalogue.htm, retrieved December 2009 (2003)
- [65] ISO 19005-1:2005, *Document management -- Electronic document file format for long-term preservation -- Part 1: Use of PDF 1.4 (PDF/A-1)*, via http://www.iso.org/iso/iso_catalogue.htm, retrieved December 2009 (2005)
- [66] ISO 28500:2009, *Information and documentation -- WARC file format*, retrieved February 2010 via http://www.iso.org/iso/iso_catalogue.htm, retrieved December 2009 (2009)
- [67] ISO 32000-1, *Document management -- Electronic document file format for long-term preservation - Part 2: Use of ISO 32000-1 (PDF/A-2)*, via http://www.iso.org/iso/iso_catalogue.htm, retrieved December 2009 (2005)
- [68] ISO/IEC 27000:2009, *Information technology - Security techniques - Information security management systems - Overview and vocabulary*,
 ISO/IEC 27001:2005, *Information technology -- Security techniques -- Information security management systems – Requirements*,
 ISO/IEC 27002:2005, *Information technology -- Security techniques -- Code of practice for information security management*,
 ISO/IEC 27003:2010, *Information technology -- Security techniques -- Information security management system implementation guidance*,

- ISO/IEC 27004:2009, *Information technology -- Security techniques -- Information security management – Measurement*,
- ISO/IEC 27005:2008, *Information technology -- Security techniques -- Information security risk management*,
- ISO/IEC 27006:2007, *Information technology -- Security techniques -- Requirements for bodies providing audit and certification of information security management systems*, via http://www.iso.org/iso/iso_catalogue.htm, retrieved April 2010 (2005-2010)
- [69] ISO/IEC 9834-8:2008, *Information technology -- Open Systems Interconnection -- Procedures for the operation of OSI Registration Authorities: Generation and registration of Universally Unique Identifiers (UUIDs) and their use as ASN.1 Object Identifier components*, via http://www.iso.org/iso/iso_catalogue.htm, retrieved December 2009 (2008)
- [70] Jacobsen, G.: *Erfaringer med høstning af det danske net 2005 – 2008 (Experiences with harvest of the Danish internet 2005 – 2008)*, In: DF Revy, vol. 31 no. 8 (2008)
- [71] Jantz, R., Giarlo, M. J.: *Digital Preservation - Architecture and Technology for Trusted Digital Repositories*, In: D-Lib Magazine, vol. 11, no. 6 (2005)
- [72] *JISC beginners Guide to digital preservation*, <http://blogs.ukoln.ac.uk/jisc-beg-dig-pres/>, retrieved August 2011
- [73] John, J. L.: *The future of saving our past*, In: Nature International Weekly Journal of Science, vol. 459, pp. 775-776 (2009)
- [74] Keeping Emulation Environments Portable, <http://www.keep-project.eu/>, retrieved August 2011
- [75] Kejser, U. B.: *Modelling the Cost and Quality of Preservation Imaging and Archiving*, Doctoral Dissertation, School of Conservation, Denmark (2009)
- [76] Kejser, U.B.: *Preservation copying of endangered historic negative collections*, In: Proceedings of the IS&T Archiving Conference, Bern, Switzerland, pp. 177-182 (2008)
- [77] Kejser, U.B., Nielsen, A.B., Thirifays, A.: *Cost Model for Digital Curation: Cost of Digital Migration*, In: Proceedings of the 6th International Conference on Preservation of Digital Objects, San Francisco, USA, pp. 98-104 (2009)
- [78] Knowledge Exchange, <http://www.knowledge-exchange.info/>, retrieved August 2011
- [79] Kopal Long-Term Digital Information Archive, <http://kopal.langzeitarchivierung.de/>, retrieved August 2011
- [80] Kounoudes, A. D., Artemi, P., Zervas, M.: *Ktisis: Building an Open Access Institutional and Cultural Repository*, In: Proceedings of the Third International Conference of Digital Heritage, pp. 504-512, Ioannides, M., Fellner, D., Georgopoulos, A., Hadjimitsis, D. G. (eds) LNCS, vol. 6436, Springer, Heidelberg (2010)
- [81] Kulovits, H., Rauber, A., Kugler, A., Brantl, M., Beinert, T., Schoger, A.: *From TIFF to JPEG 2000? Preservation Planning at the Bavarian State Library Using a Collection of Digitized 16th Century Printings*, In: D-Lib Magazine, vol. 15, no. 11/12 (2009)
- [82] Kuny, T.: *A Digital Dark Ages? Challenges in the Preservation of Electronic Information*, IFLANET International Preservation News, no. 17, <http://archive.ifla.org/IV/ifla63/63kuny1.pdf>, retrieved August 2011 (1998)
- [83] Kunze, J.: *The ARK Persistent Identifier Scheme*, <http://tools.ietf.org/html/draft-kunze-ark-14.txt>, retrieved August 2011 (2007)

- [84] Kunze, J., Littman, J., Vargas, B.: *The BagIt File Packaging Format (V0.97)*, <http://www.ietf.org/internet-drafts/draft-kunze-bagit-06.txt>, retrieved August 2011 (2011)
- [85] Lavoie, B.: *PREMIS With a Fresh Coat of Paint*, In: D-Lib Magazine, vol. 14 no. 5/6 (2008)
- [86] Lavoie, B., Dempsey, L.: *Thirteen Ways of Looking at...Digital Preservation*, In: D-Lib Magazine vol. 10 no. 7/8 (2004)
- [87] Leach, P., Salz, R.: *A Universally Unique Identifier (UUID) URN Namespace*, <http://tools.ietf.org/rfc/rfc4122.txt>, retrieved August 2011
- [88] Lee, K., Slattery, O., Lu, R., Tang, X., McCrary, V.: *The State of the Art and Practice in Digital Preservation*, In: Journal of Research of the National Institute of Standards and Technology, vol. 107, no. 1, pp. 93-106 (2002)
- [89] Library of Congress, <http://www.loc.gov/>, retrieved August 2011
- [90] *Library of Congress and DuraCloud Launch Pilot Program Using Cloud Technologies to Test Perpetual Access to Digital Content*, News from the Library of Congress, July 14 2009, <http://www.loc.gov/today/pr/2009/09-140.html>, retrieved August 2011 (2009)
- [91] *Local Digital Format Registry (LDFR), File Format Guidelines for Preservation and Long-term Access Version 1.0*, Library and Archives Canada, <http://www.collectionscanada.gc.ca/digital-initiatives/012018-2210-e.html>, retrieved August 2011
- [92] Long, A. S.: *Long-term Preservation of Web Archives – Experimenting with Emulation and Migration Methodologies*, IIPC report, http://netpreserve.org/publications/NLA_2009_IIPC_Report.pdf, retrieved August 2011 (2009)
- [93] Lynch, C.: *Integrity Issues in Electronic Publishing*, Peek, R. P., Newby, G. B. (eds.) *Scholarly Publishing: The Electronic Frontier*, Cambridge: The MIT Press, pp. 133-145 (1996)
- [94] Lynch, C.: *The Integrity of Digital Information: Mechanics and Definitional Issues*, In: Journal of the American Society for Information Science, vol. 45, issue 10, pp. 737-744 (1994)
- [95] Manes, S.: *Time and Technology Threaten Digital Archives ...*, The New York Times, April 1998, <http://www.nytimes.com/1998/04/07/science/time-and-technology-threaten-digital-archives.html>, retrieved August 2011
- [96] Maniatis, P., Roussopoulos, M., Giuli, T. J., Rosenthal, D. S. H., Baker, M.: *The LOCKSS Peer-to-Peer Digital Preservation System*, In: ACM Transactions on Computer Systems, vol. 23, no. 1, pp. 2-50 (2005)
- [97] Manson, P.: *Digital Preservation Research: An Involving Landscape*, In: Proceedings of the 7th International Conference on Preservation of Digital Objects, Vienna, Austria, pp. 18 (2010)
- [98] Marcum, D.: *Introduction: The Changing Preservation Landscape*, In: Proceedings of The State of Digital Preservation: An International Perspective, Washington, D.C., USA (2002)
- [99] McDonough, J., Olendorf, R., Kirchenbaum, M., Kraus, K., Reside, D., Donahue, R., Phelps, A., Egert, C., Lowood, H., Rojo, S.: *Preserving Virtual Worlds - Final Report*, <http://hdl.handle.net/2142/17097>, retrieved August 2011 (2010)
- [100] MetaArchive, <http://www.metaarchive.org/>, retrieved August 2011
- [101] *Metadata Encoding & Transmission Standard*, www.loc.gov/standards/mets/mets-home.html, retrieved August 2011
- [102] METS, <http://www.loc.gov/standards/mets/>, retrieved August 2011
- [103] *Micro-Services*, <https://www.irods.org/index.php/Micro-Services>, retrieved August 2011

- [104] Miller, E.: *An Introduction to the Resource Description Framework*, In: D-Lib Magazine, May 1998 (1998)
- [105] Minor, D., Phillips, M., Schultz, M.: *Chronopolis and MetaArchive: Preservation Cooperation*, In: Proceedings of the 7th International Conference on Preservation of Digital Objects, Vienna, Austria, pp. 249-254 (2010)
- [106] Minor, D., Sutton, D., Kozbial, A., Burek, M., Smorul, M.: *Chronopolis: Preserving our Digital Heritage*, In: Proceedings of the 6th International Conference on Preservation of Digital Objects, San Francisco, USA, pp. 141-147 (2009)
- [107] Minor, D., Sutton, D., Kozbial, A., Westbrook, B., Burek, M., Smorul, M.: *Chronopolis Digital Preservation Network*, In: The International Journal of Digital Curation, vol. 5, issue 1 (2010)
- [108] MIX, <http://www.loc.gov/standards/mix/>, retrieved August 2011
- [109] MODS, <http://www.loc.gov/standards/mods/>, retrieved August 2011
- [110] Moore, R. W., MacKenzie, S.: *Automated Validation of Trusted Digital Repository Assessment Criteria*, In: Journal of Digital Information, vol. 8, no. 2 (2007)
- [111] National Archives of the United Kingdom, <http://www.nationalarchives.gov.uk/default.htm>, retrieved August 2011
- [112] National Digital Stewardship Alliance, <http://www.digitalpreservation.gov/ndsaa/>, retrieved August 2011
- [113] National Library of Norway: *Digitization of books in the National Library - methodology and lessons learned*, http://www.nb.no/content/download/2326/18198/version/1/file/digitizing-books_sep07.pdf via <http://www.nb.no/english/facts/about-the-national-library>, retrieved August 2011 (2007)
- [114] National Science Foundation, <http://www.nsf.gov/about/>, retrieved August 2011
- [115] nestor (German Network of Expertise in long-term storage of digital resources): *Kriterienkatalog vertrauenswürdige digitale Langzeitarchive, Version 2*, via <http://www.langzeitarchivierung.de/arbeitsgruppen/agkritkat.htm>, retrieved August 2011 (2008)
- [116] Netarchive.dk, www.netarkivet.dk, retrieved August 2011
- [117] NetarchiveSuite, <http://netarchive.dk/suite/>, retrieved August 2011
- [118] Netherlands Coalition for Digital Preservation, <http://www.ncdd.nl/en/index.php>, retrieved August 2011
- [119] Nielsen, E. K.: *Den danske Nationallitteratur før 1700 digitaliseres - igennem et banebrydende internationalt samarbejde (The Danish National Literature before 1700 is being digitised - through path-breaking international collaboration)*, <http://www.kb.dk/da/materialer/kulturarv/Nyhed2010.html>, retrieved August 2011 (2010)
- [120] NIST/Library of Congress (LoC) *Optical Disc Longevity Testing Procedure*, NIST Special Publication 500-263 (2005)
- [121] Norcen, R., Podesser, M., Pommer, A., Schmidt, H. P., Uhl, A.: *Confidential storage and transmission of medical image data*, In: Journal of Computers in Biology and Medicine, vol. 33, issue 3, pp. 277-292 (2003)
- [122] OCLC (Online Computer Library Center) & CRL (Center for Research Libraries): *Trustworthy Repositories Audit & Certification (TRAC): Criteria and Checklist, Version 1.0*,

- http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf, retrieved August 2011 (2007)
- [123] Oltmans, E., van Diessen, R. J., van Wijngaarden, H.: *Preservation Functionality in a Digital Archive*, In: Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries, Tucson, Arizona, USA (2004)
- [124] Open Planets Foundation, <http://www.openplanetsfoundation.org/>, retrieved August 2011
- [125] *Optical media longevity – the X-lab*, <http://www.thexlab.com/faqs/opticalmedialongevity.html>, retrieved August 2011
- [126] Patterson, D. A., Gibson, G., Katz, R. H.: *A case for redundant arrays of inexpensive disks (RAID)*, In: Proceedings of the 1988 ACM SIGMOD international conference on Management of data, Chicago, Illinois, United States, pp. 109-116 (1988)
- [127] Paul, A., Hagmann, J.: *Challenges of Long-Term Archiving in the Pharmaceutical Industry*, In: Proceedings of the IS&T Archiving Conference, Bern, Switzerland, pp. 26-29 (2008)
- [128] Planets - Preservation and Long-term Access through NETWORKED Services, <http://www.planets-project.eu/> or <http://www.openplanetsfoundation.org/>, retrieved August 2011
- [129] Pligtafleivering (legal deposit in Denmark), <http://www.pligtafleivering.dk/index.htm>, retrieved August 2011
- [130] *Portico's Certification CRL Audit* <http://www.portico.org/digital-preservation/the-archive-content-access/archive-certification/>, retrieved August 2011 (2010)
- [131] PREMIS, <http://www.loc.gov/standards/premis/>, retrieved August 2011
- [132] *PREMIS Data Dictionary for Preservation Metadata version 2.0*, <http://www.loc.gov/standards/premis/v2/premis-2-0.pdf>, retrieved August 2011, (2008)
- [133] Preserving Access to Digital Information, <http://www.nla.gov.au/padi/>, retrieved August 2011
- [134] *Private LOCKSS Networks*, http://lockss.stanford.edu/lockss/Private_LOCKSS_Networks, retrieved August 2011
- [135] Pyrounakis, G., Saidis, K., Nikolaidou, M., Lourdi, I.: *Designing an Integrated Digital Library Framework to Support Multiple Heterogeneous Collections*, In: Proceedings of the 8th European Conference on Research and Advanced Technology for Digital Libraries, pp. 26-37, Heery, R., Lyon, L. (eds.) LNCS vol. 3232, Springer, Heidelberg (2004)
- [136] Rauch, C., Rauber, A.: *Preserving Digital Media: Towards a Preservation Solution Evaluation Metric*, In: Proceedings of the 7th International Conference on Asian Digital Libraries, pp. 19-32, Chen, Z., Chen, H., Miao, Q., Fu, Y., Fox, E., Lim, E. (eds.) LNCS vol. 3334, Springer, Heidelberg (2005)
- [137] Razum, M., Schwichtenberg, F., Wagner, S., Hoppe, M.: *ESciDoc infrastructure: a Fedora-based e-research framework*, In: Proceedings of the 13th European Conference on Research and Advanced Technology for Digital Libraries, pp. 227-238, Agosti, M., Borbinha, J., Kapidakis, S., Papatheodorou, C., Tsakonas, G. (eds.) LNCS, vol. 5714, Springer, Heidelberg (2009)
- [138] Reich, V., Rosenthal, D. S. H.: *Distributed Digital Preservation: Private LOCKSS Networks as Business, Social, and Technical Frameworks*, In: Library Trends, vol. 57, no. 3, pp. 461-475 (2009)
- [139] Reich, V., Rosenthal, D. S. H.: *LOCKSS: A Permanent Web Publishing and Access System*, In: D-Lib Magazine, vol. 7, no. 6 (2001)

- [140] Reich, V., Rosenthal, D. S. H.: *LOCKSS (Lots of copies Keep stuff safe)*, *The New Review of Academic Librarianship*, vol. 6, no. 1, pp. 155-161 (2000)
- [141] Riva, P.: *Functional requirements for bibliographic records: Introducing the Functional Requirements for Bibliographic Records and related IFLA developments*, In: *Bulletin of the American Society for Information Science and Technology*, vol. 33, issue 6 (2008)
- [142] RLG-OCLC (Research Libraries Group - Online Computer Library Center): *Trusted Digital Repositories: Attributes and Responsibilities*, an RLG-OCLC report, <http://www.oclc.org/research/activities/past/rlg/trustedrep/repositories.pdf>, retrieved August 2011 (2002)
- [143] Rosenthal, D. S. H.: *Bit Preservation: A Solved Problem?*, In: *The International Journal of Digital Curation*, vol. 5, no. 1 (2010)
- [144] Rosenthal, D. S. H., Robertson, T., Lipkis, T., Reich, V., Morabito, S.: *Requirements for Digital Preservation Systems, A Bottom-Up Approach*, In: *D-Lib Magazine*, vol. 11, no. 11 (2005)
- [145] Rosenzweig, R.: *Scarcity or Abundance? Preserving the Past in a Digital Era*, <http://chnm.gmu.edu/digitalhistory/links/pdf/introduction/0.6b.pdf>, retrieved August 2011
- [146] Ross, S., Hedstrom, M.: *Preservation research and sustainable digital libraries*, In: *The International Journal of Digital Libraries*, vol. 5, no. 4, pp. 317-324, Springer, Heidelberg (2005)
- [147] Rothenberg, J.: *Ensuring the longevity of digital information*, In: *Scientific American*, vol. 272, no. 1, pp. 42-47, extended version from 1998 retrieved August 2011 from <https://www.clir.org/pubs/archives/ensuring.pdf> (1995)
- [148] Roussopoulos M., Baker M., Rosenthal D. S. H., Giuli T. J.: *2 P2P or Not 2 P2P?*, In: *Proceedings of Third International Workshop on Peer-to-Peer Systems*, pp. 33-43, Voelker, G. M., Shenker, S. (eds), vol. 3279 LNCS, Springer, Heidelberg (2005)
- [149] Roussopoulos, M., Bungale, P.: *Stealth modification versus nuisance attacks in the LOCKSS peer-to-peer digital preservation system*, In: *Journal of Peer-to-Peer Networking and Applications*, vol. 3, no. 4, pp. 265-276, Springer, Heidelberg (2010)
- [150] Rusbridge, C.: *Excuse me...some digital preservation fallacies?*, In: *Ariadne Web Magazine*, issue 46 (2006)
- [151] Rusbridge, A., Ross, S.: *The UK LOCKSS Pilot Programme: A Perspective from the LOCKSS Technical Support Service*, In: *International Journal of Digital Curation*, vol. 2, no. 2 (2007)
- [152] Saraiva, J. S., Silva, A. R.: *Design Issues for an Extensible CMS-Based Document Management System*, In: *Revised selected papers from First International Joint Conference*, Funchal, Madeira, Portugal, October 2009, pp. 323-336, Fred, A., Dietz, J. L. G., Liu, K., Filipe J. (eds.) *Communications in Computer and Information Science*, vol. 128, part 4, Springer, Heidelberg (2011)
- [153] Saroiu, S., Gummadi, P. K., Gribble, S. D.: *A Measurement Study of Peer-to-Peer File Sharing Systems*, Technical Report UW-CSE-01-06-02, Department of Computer Science & Engineering, University of Washington, <http://www.cs.ucsb.edu/~almeroth/classes/F05.276/papers/p2p-measure.pdf>, Retrieved August 2011 (2002)
- [154] Saur, K.G.: *The Relative Stabilities of Optical Disc Formats*
In: *Restaurator - International Journal for the Preservation of Library and Archival Material*, vol. 26, no. 2 (2005)

- [155] Schmitz, D.: *The Seamless Cyberinfrastructure: The Challenges of Studying Users of Mass Digitization and Institutional Repositories*, <http://www.clir.org/pubs/archives/schmitz.pdf>, retrieved August 2011
- [156] Schroeder, B., Gibson, G. A.: *Disk failures in the real world: What does an MTTF of 1,000,000 hours mean to you?*, In: Proceedings of 5th USENIX Conference on File and Storage Technologies, pp. 1-16 (2007)
- [157] Seadle, M.: *Archiving in the networked world: LOCKSS and national hosting*, In: Library Hi Tech, vol. 28, Issue 4, pp. 710-717 (2010)
- [158] Skinner, K., Halbert, M.: *The MetaArchive Cooperative: A Collaborative Approach to Distributed Digital Preservation*, In: Library Trends vol. 57, no. 3, pp. 371-392 (2009)
- [159] Slattery, O., Lu, R., Zheng, J., Byers, F., Tang, X.: *Stability Comparison of Recordable Optical Discs—A Study of Error Rates in Harsh Conditions*, In: Journal of Research of the National Institute of Standards and Technology, vol. 109, no. 5 (2004)
- [160] Smith, M., Bass, M., McClellan, G., Tansley, R., et al.: *DSpace : An Open Source Dynamic Digital Repository*, In: D-Lib Magazine, vol. 9, no. 1 (2003)
- [161] *SNIA Green Storage Initiative*, <http://www.snia.org/forums/green/>, retrieved August 2011
- [162] Sollins, K., Masinter, L.: *Functional Requirements for Uniform Resource Names*, <http://tools.ietf.org/html/rfc1737>, retrieved August 2011 (1994)
- [163] Stephen, A., Kunze, J., Loy, D.: *An Emergent Micro-Services Approach to Digital Curation Infrastructure*, In: International Journal of Digital Curation, vol. 5, issue 1, pp. 172-186 (2010)
- [164] Stock, C., Rocklin, E., Cordier, A.: *LARA—Open access to scientific and technical reports*, In: Publishing Research Quarterly, vol. 22, no. 1, pp. 42-51 (2006)
- [165] Storer, M. W., Greenan, K., Miller, E. L.: *Long-Term Threats to Secure Archives*, In: Proceedings of the second ACM workshop on Storage security and survivability, Alexandria, Virginia, USA, pp. 9-16 (2006)
- [166] Strodl, S., Petrov, P., Greifeneder, M., Rauber, A.: *Automating Logical Preservation for Small Institutions with Hoppla*, In: Proceedings of the 14th European Conference on Research and Advanced Technology for Digital Libraries, pp. 124-135, Lalmas, M., Jose, J., Rauber, A., Sebastiani, F., Frommholz, I. (eds.) LNCS, vol. 6273, Springer, Heidelberg (2010)
- [167] Strodl, S., Petrov, P., Rauber, A.: *Research on Digital Preservation within projects co-funded by the European Union in the ICT programme*, http://cordis.europa.eu/fp7/ict/telearn-digicult/report-research-digital-preservation_en.pdf, retrieved August 2011 (2011)
- [168] Tansley, R., Smith, M., Walker, J. H.: *The DSpace Open Source Digital Asset Management System: Challenges and Opportunities*, In: Proceedings of the 9th European Conference on Research and Advanced Technology for Digital Libraries, pp. 242–253, Rauber, A., Christodoulakis, S., Tjoa, A. M. (eds.) LNCS vol. 3652, Springer, Heidelberg (2005)
- [169] Text Encoding Initiative, <http://www.tei-c.org/index.xml>, retrieved August 2011
- [170] *Text Encoding Initiation (P4)*, <http://www.tei-c.org/Guidelines/P4/>, retrieved August 2011
- [171] The Handle System®, <http://www.handle.net>, retrieved August 2011
- [172] Thaller, M.: *The eXtensible Characterisation Languages - XCL*, Verlag Dr. Kovač, Hamburg, 2009
- [173] *The Future of the Past – Shaping new visions for EU-research in digital preservation*, Workshop, 4-5 May 2011, Luxembourg,

http://cordis.europa.eu/fp7/ict/telearn-digicult/digicult-future-digital-preservation_en.html,
retrieved August 2011 (2011)

- [174] *The Murray Research Archive's policy for digital archiving*,
<http://www.murray.harvard.edu/policies>, retrieved August 2011
- [175] *The Preservation Management of Digital Material Handbook*,
http://www.dpconline.org/component/docman/doc_download/299-digital-preservation-handbook, retrieved August 2011 (2008)
- [176] The technical registry PRONOM,
<http://www.nationalarchives.gov.uk/aboutapps/pronom/default.htm>, retrieved August 2011
- [177] Thibodeau, K.: *Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years*, In: Proceedings of The State of Digital Preservation: An International Perspective, Washington, D.C., USA (2002)
- [178] Tonkin, E.: *Persistent Identifiers: Considering the Options*, In: Adriadne magazine, issue 56, July 2008 (2008)
- [179] *Understanding CD-R & CD-RW Disc Longevity*, <http://www.osta.org/technology/cdqa13.htm>,
retrieved August 2011
- [180] U.S. Food and Drug Administration, <http://www.fda.gov/>, retrieved August 2011
- [181] van der Hoeven, J., Lohman B., Verdegem, R.: *Requirements for Applying Emulation as a Preservation Strategy*, In: Proceedings of the IS&T Archiving Conference, Bern, Switzerland, pp. 71-76 (2008)
- [182] van Garderen, P.: *ARCHIVEMATICA: Using Micro-Services and Open-Source Software to Deliver a Comprehensive Digital Curation Solution*, In: Proceedings of the 7th International Conference on Preservation of Digital Objects, Vienna, Austria, pp. 145-149 (2010)
- [183] van Wijngaarden, H., Rog, J., Marijnen, P.: *Building Blocks for the new KB E-Depot*, In: Proceedings of the 7th International Conference on Preservation of Digital Objects, Vienna, Austria, pp. 315-320 (2010)
- [184] Vermaaten, S.: *A Checklist and a Case for Documenting PREMIS-METS Decisions in a METS Profile*, In: D-Lib Magazine, vol. 16, no. 9/10 (2010)
- [185] Vienna University of Technology, Institute of Software Technology and Interactive Systems, Digital Preservation, <http://www.ifs.tuwien.ac.at/dp/>, retrieved August 2011
- [186] Wan, M., Moore, R., Rajasekar, A.: *Integration of Cloud Storage with Data Grids*, Presented at: Third International Conference on the Virtual Computing Initiative, North Carolina, USA,
https://www.irods.org/pubs/DICE_icvci3_mainpaper_pub-0910.pdf, retrieved August 2011 (2009)
- [187] Wassen, C. S.: *System Analysis, Design and Development Concepts, Principles and Practices*, John Wiley & Sons, Inc., Hoboken, New Jersey (2006)
- [188] Waters, D., Garrett, J.: *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information*, <http://www.oclc.org/research/activities/past/rlg/digpresstudy/final-report.pdf>,
retrieved August 2011 (1996)
- [189] *Webster's Encyclopedic Unabridged Dictionary of the English Language*, Stein Jess (eds.) Random House Inc. (1996)
- [190] Wilson, A.: *Significant Properties Report, InSPECT Work Package 2.2 Draft/Version: V2*,
http://www.significantproperties.org.uk/wp22_significant_properties.pdf, retrieved August 2011 (2007)

- [191] World of Warcraft, <http://www.worldofwarcraft.com/>, retrieved August 2011
- [192] Wright, R., Miller, A., Addis, M.: *The Significance of Storage in the "Cost of Risk" of Digital Preservation*, In: The International Journal of Digital Curation, vol. 4, issue 3, pp. 105-122 (2009)
- [193] Wu, L., Buyya, R.: *Service Level Agreement (SLA) in Utility Computing Systems*, Technical Report from The University of Melbourne of Australia, <http://arxiv.org/ftp/arxiv/papers/1010/1010.2881.pdf>, retrieved August 2011 (2010)
- [194] Zierau, E., van Wijk, C.: *The Planets Approach to Migration Tools*, In: Proceedings of the IS&T Archiving Conference, Bern, Switzerland, pp. 30-35 (2008)

Papers

Paper A. Representation of Digital Material preserved in a Library Context

The paper reference is:

Zierau, E.: *Representation of Digital Material preserved in a Library Context*, In: Proceedings of the 7th International Conference on Preservation of Digital Objects, Vienna, Austria, pp. 329-337, Copyrights held by Oesterreichische computer gesellschaft, Printed by Börse Druck, www.boersedruck.at, ISBN 978-3-85403-262-5 (2010)

Full peer-reviewed paper

Please notice that the paper is under iPRES copyright restrictions.

REPRESENTATION OF DIGITAL MATERIAL PRESERVED IN A LIBRARY CONTEXT

Eld Zierau

The Royal Library of Denmark
Dep. of Digital Preservation
P.O.BOX 2149
1016 Copenhagen K, Denmark

Under copyright, but available from:

<http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/zierau-19.pdf>

Under copyright, but available from:

<http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/zierau-19.pdf>

Under copyright, but available from:

<http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/zierau-19.pdf>

Under copyright, but available from:

<http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/zierau-19.pdf>

Under copyright, but available from:

<http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/zierau-19.pdf>

Under copyright, but available from:

<http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/zierau-19.pdf>

Under copyright, but available from:

<http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/zierau-19.pdf>

Under copyright, but available from:

<http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/zierau-19.pdf>

Paper B. Preservation of Digitised Books in a Library Context

The paper reference is:

Zierau, E., Jensen, C.: *Preservation of Digitised Books in a Library Context*, In: Proceedings of the 7th International Conference on Preservation of Digital Objects, Vienna, Austria, pp. 61-69, Copyrights held by Oesterreichische computer gesellschaft, Printed by Börse Druck, www.boersedruck.at, ISBN 978-3-85403-262-5 (2010)
Full peer-reviewed paper

Please notice that the paper is under iPRES copyright restrictions.

Known typos and errors in the paper

- *Second paragraph after table 4:*

The sentence

"Most of the JPEGs have errors in letter recognition. Especially book {a} has many errors in the SC1 JPEG"

can be misleading, since the errors were found in samples with fewer number of pages.

PRESERVATION OF DIGITISED BOOKS IN A LIBRARY CONTEXT

Eld Zierau

Claus Jensen

The Royal Library of Denmark
Dep. of Digital Preservation
P.O.BOX 2149
1016 Copenhagen K, Denmark

Under copyright, but available from:

<http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/zierau-18.pdf>

Under copyright, but available from:

<http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/zierau-18.pdf>

Under copyright, but available from:

<http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/zierau-18.pdf>

Under copyright, but available from:

<http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/zierau-18.pdf>

Under copyright, but available from:

<http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/zierau-18.pdf>

Under copyright, but available from:

<http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/zierau-18.pdf>

Under copyright, but available from:

<http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/zierau-18.pdf>

Under copyright, but available from:

<http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/zierau-18.pdf>

Paper C. Archive Design Based on Planets Inspired Logical Object Model

The paper reference is:

Zierau, E., Johansen, A.S.: *Archive Design Based on Planets Inspired Logical Object Model*, In: Proceedings of the 12th European Conference on Research and Advanced Technology for Digital Libraries, pp. 37-40, Christensen-Dalsgaard, B., Castelli, D., Jurik, B.A., Lippincott, J. (eds.) LNCS, vol. 5173, Publisher: Springer-Verlag Berlin Heidelberg, ISBN 978-3-540-87598 (2008)
Short peer- reviewed paper

Please notice that the paper is under Springer (ECDL) copyright restrictions.

Known typos and errors in the paper

- *The third bullet in first bullet in section 3:*
The term transformation should be migration in order only to use consequent use of one term.

Archive Design based on Planets inspired Logical Object Model

Eld Zierau, Anders Sewerin Johansen

The Royal Library of Denmark, P.O.BOX 2149, 1016 Copenhagen K, Denmark
elzi@kb.dk, asjo@kb.dk

Under copyright

In: Proceedings of the 12th European Conference on Research and Advanced Technology for Digital Libraries, pp. 37-40, Christensen-Dalsgaard, B., Castelli, D., Jurik, B.A., Lippincott, J. (eds.) LNCS, vol. 5173, Publisher: Springer-Verlag Berlin Heidelberg, ISBN 978-3-540-87598 (2008)

Under copyright

In: Proceedings of the 12th European Conference on Research and Advanced Technology for Digital Libraries, pp. 37-40, Christensen-Dalsgaard, B., Castelli, D., Jurik, B.A., Lippincott, J. (eds.) LNCS, vol. 5173, Publisher: Springer-Verlag Berlin Heidelberg, ISBN 978-3-540-87598 (2008)

Under copyright

In: Proceedings of the 12th European Conference on Research and Advanced Technology for Digital Libraries, pp. 37-40, Christensen-Dalsgaard, B., Castelli, D., Jurik, B.A., Lippincott, J. (eds.) LNCS, vol. 5173, Publisher: Springer-Verlag Berlin Heidelberg, ISBN 978-3-540-87598 (2008)

Under copyright

In: Proceedings of the 12th European Conference on Research and Advanced Technology for Digital Libraries, pp. 37-40, Christensen-Dalsgaard, B., Castelli, D., Jurik, B.A., Lippincott, J. (eds.) LNCS, vol. 5173, Publisher: Springer-Verlag Berlin Heidelberg, ISBN 978-3-540-87598 (2008)

Paper D. Cross Institutional Cooperation on a Shared Bit Repository

The paper reference is:

Zierau, E., Kejser, U.B.: *Cross Institutional Cooperation on a Shared Bit Repository*,
In: Journal of the World Digital Libraries, vol. 3, issue 1, pp. 11-21, Publisher: TERI Press,
New Delhi, ISSN 0974-567-X (2010)
(Awarded with *best paper award – international* at *International Conference on Digital
Libraries*, New Delhi, India, 2010)
Full peer-reviewed paper

Please notice that the paper is under WDL and ICDL copyright restrictions.

Known typos and errors in the paper

- *A few lines before figure 4:*
The sentence
"R-Ingest & BR-Ingest and IR-Access & BR-Ingest"
should instead be
"IR-Ingest & BR-Ingest and IR-Access & BR-Access".
- *Last sentence in 2nd paragraph in introduction:*
The sentence
"The business model for the bit repository will be based on pay per service"
can be misleading, since it can be very hard to define what a service is in a bit repository.

Cross Institutional Cooperation on a Shared Bit Repository

Eld Zierau

The Royal Library of Denmark, P.O.BOX 2149, DK-1016 Copenhagen K, elzi@kb.dk

Ulla Bøgvad Kejser

School of Conservation, Esplanaden 34, DK-1263 Copenhagen K, ubk@kb.dk

Under copyright

In: Journal of the World Digital Libraries, vol. 3, issue 1, pp. 11-21,
Publisher: TERI Press

Or

In: *Proceedings of the International Conference on Digital Libraries*,
New Delhi, India, 2010

Under copyright

In: Journal of the World Digital Libraries, vol. 3, issue 1, pp. 11-21,
Publisher: TERI Press

Or

In: *Proceedings of the International Conference on Digital Libraries*,
New Delhi, India, 2010

Under copyright

In: Journal of the World Digital Libraries, vol. 3, issue 1, pp. 11-21,
Publisher: TERI Press

Or

In: *Proceedings of the International Conference on Digital Libraries*,
New Delhi, India, 2010

Under copyright

In: Journal of the World Digital Libraries, vol. 3, issue 1, pp. 11-21,
Publisher: TERI Press

Or

In: *Proceedings of the International Conference on Digital Libraries*,
New Delhi, India, 2010

Under copyright

In: Journal of the World Digital Libraries, vol. 3, issue 1, pp. 11-21,
Publisher: TERI Press

Or

In: *Proceedings of the International Conference on Digital Libraries*,
New Delhi, India, 2010

Under copyright

In: Journal of the World Digital Libraries, vol. 3, issue 1, pp. 11-21,
Publisher: TERI Press

Or

In: *Proceedings of the International Conference on Digital Libraries*,
New Delhi, India, 2010

Under copyright

In: Journal of the World Digital Libraries, vol. 3, issue 1, pp. 11-21,
Publisher: TERI Press

Or

In: *Proceedings of the International Conference on Digital Libraries*,
New Delhi, India, 2010

Under copyright

In: Journal of the World Digital Libraries, vol. 3, issue 1, pp. 11-21,
Publisher: TERI Press

Or

In: *Proceedings of the International Conference on Digital Libraries*,
New Delhi, India, 2010

Under copyright

In: Journal of the World Digital Libraries, vol. 3, issue 1, pp. 11-21,
Publisher: TERI Press

Or

In: *Proceedings of the International Conference on Digital Libraries*,
New Delhi, India, 2010

Under copyright

In: Journal of the World Digital Libraries, vol. 3, issue 1, pp. 11-21,
Publisher: TERI Press

Or

In: *Proceedings of the International Conference on Digital Libraries*,
New Delhi, India, 2010

Paper E. Evaluation of Bit preservation Strategies

The paper reference is:

Zierau, E., Kejser, U.B., Kulovits, H.: *Evaluation of Bit Preservation Strategies*, In: Proceedings of the 7th International Conference on Preservation of Digital Objects, Vienna, Austria, pp. 161-169, Copyrights held by Oesterreichische computer gesellschaft, Printed by Börse Druck, www.boersedruck.at, ISBN 978-3-85403-262-5 (2010)
Full peer-reviewed paper

Please notice that the paper is under iPRES copyright restrictions.

Known typos and errors in the paper

- *In the section just after table 7:*
S4 should be replaced with S3.
- *In the Table 7. Plato results for SLA cases to M1*
The table shows values where confidentiality is given weight 80% and integrity is given weight 20% instead of a 50% - 50% distribution. The conclusions are however the same since there is only one SLA that is not ruled out. The correct totals are given in Table 6 in Appendix I.
- *In the Table 8. Plato results for SLA cases to M2*
The table shows values where confidentiality is given weight 20% and integrity is given weight 80% instead of a 50% - 50% distribution. The conclusions are however the same, since the requirement to all confidentiality requirements is low, which means that they for all SLAs are given maximum utility function. Thus for all SLAs the confidentiality is given the same contribution, and variations on integrity are the same no matter that the percentage of the results are 80% instead of 50%. The correct totals are given in Table 9 in Appendix I.

EVALUATION OF BIT PRESERVATION STRATEGIES

Eld Zierau & Ulla Bøgvad Kejser

The Royal Library of Denmark
Dep. of Digital Preservation
P.O.BOX 2149
1016 Copenhagen K, Denmark

Hannes Kulovits

Vienna University of Technology
Inst. of SW Tech. & Interactive Sys.
Favoritenstraße 9-11/188/2
A-1040 Wien, Austria

Under copyright, but available from:

<http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/zierau-31.pdf>

Under copyright, but available from:

<http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/zierau-31.pdf>

Under copyright, but available from:

<http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/zierau-31.pdf>

Under copyright, but available from:

<http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/zierau-31.pdf>

Under copyright, but available from:

<http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/zierau-31.pdf>

Under copyright, but available from:

<http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/zierau-31.pdf>

Under copyright, but available from:

<http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/zierau-31.pdf>

Under copyright, but available from:

<http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/zierau-31.pdf>

Appendices

Appendix I. Detailed Calculations using Plato

This appendix provides the detailed calculations behind the results in section 5 in Paper E. Note that the results listed are made on relatively simple cases, and should therefore only be seen as documentation for how the evaluation methodology works.

Section III.1 gives a short introduction to the specific use of Plato. Section III.2 gives the general results of the SLA cases in Paper E without relation to specific digital material requirements. Section III.3 gives the specific results for the different digital material cases described in Paper E.

Table of contents of the appendix:

I.1. Using Plato.....	165
I.2. General BR results	166
I.3. Material specific results.....	168
I.3.1 Plato results for digitally born diaries	168
I.3.2 Plato results for digitally born images.....	170
I.3.3 Plato results for digitised books.....	171

I.1 Using Plato

Plato is a planning tool primarily used for planning of functional preservation actions. The Planets preservation planning workflow consists of three main stages:

- *Requirements definition*

This includes specification of requirements in a quantifiable way, starting at high-level objectives and breaking them down into measurable criteria, thus creating an objective tree which forms the basis of the evaluation of alternative strategies. The requirements tree with measurable requirements, as illustrated in Figure 4 in Paper E, is specified in the open source mind mapping program Freemind (http://freemind.sourceforge.net/wiki/index.php/Main_Page).

- *Evaluation of potential strategies*

In traditional use of Plato the evaluation is made on which tools to use in preservation actions.

In the bit preservation strategy this instead involves evaluation of potential uses of a concrete bit repository expressed in the service level agreements.

This means the measurable requirements from the requirements tree are evaluated against how well the specific material requirements are met by the different service level agreements.

The result is based on aggregation of the leaf values in the requirements tree

- *Analysis of the results*

Analysis of the results is based on output from the evaluation which includes two types of aggregation (taken from <http://olymp.ifs.tuwien.ac.at:8080/plato/help/aggregation.html>):

- *Multiplication*

The first step is to raise the comparable value per leaf to the weight of that leaf. The results are then multiplied throughout the tree for the whole alternative. The final ranking is based on a rational scale. The multiplication method highlights alternatives with drop out values, as these alternatives with leaf values zero have a final root value of zero.

- *Sum*

The comparable values are multiplied by their weights. These values are summed up to a single comparable value per alternative. The sum method offers a final ranking on a rational scale. Leaf values that score zero (drop-out value) have no decisive effect on the final root value.

The presented results only include the sum aggregation, since the cases do not include weight of leaves.

I.2 General BR results

The general results for the different SLA cases are given in Table 1. The results listed are the same as the ones given in table 4 in Paper E. The difference is that there are extra notes for the SLAs in order to make it more readable, and the requirements are spelled out.

Requirement	SLA case					
	S1 DK	S2 Full replica -> checksum	S3 For confi- dential material	S4 Full re- plica on cloud	S5 For bit safe material	S6 As S5 with extra checksum
Confidentiality						
Authorisation security violation	Medium	Medium	High	Low	Low	Low
Physical security violation	Medium	Medium	High	Low	Medium	Medium
Technical security violation	Medium	Medium	High	Low	Medium	Medium
Transmission security violation	Medium	Medium	High	Low	Medium	Medium
Integrity						
<i>audit frequency</i>						
Bit errors are found	Medium	Medium	Medium	Medium	Medium	High
Bit errors are corrected in time	Medium	Low	Low	Low	Medium	Medium
<i>Independence -technical</i>						
Different hardware/media	Medium	Medium	High	Low	Medium	Medium
Different operating system	High	High	High	Low	High	High
Different software	Medium	High	High	Low	High	High
<i>Independence -organisation</i>						
Different internal damage preventions	Medium	Medium	Medium	Low	Medium	Medium
Different war/terror attacks damage prev.	Medium	Medium	Low	Low	High	High
Different virus, worms attacks damage prev.	Medium	Medium	High	Low	Medium	Medium
Different natural disaster damage prev.	Medium	Medium	Medium	Low	High	High

Table 1 BR-ReMS results of requirement fulfilment.

Evaluation of the different SLA's ability to meet different requirements can be done *without* relating them to specific material requirements. This is done by applying the uniform scale with values from 0-5 which Plato uses for values to nodes in the requirements as input for the utility analysis. The values given in this example are:

- Low = 1
- Medium = 3
- High = 5

This gives the values listed in Table 2.

Requirement	SLA case					
	S1 DK	S2 Full replica -> checksum	S3 For confi- dential material	S4 Full replica on cloud	S5 For bit safe material	S6 As S5 with extra checksum
Confidentiality						
Authorisation security violation	3	3	5	1	1	1
Physical security violation	3	3	5	1	3	3
Technical security violation	3	3	5	1	3	3
Transmission security violation	3	3	5	1	3	3
Integrity						
<i>audit frequency</i>						
Bit errors are found	3	3	3	3	3	5
Bit errors are corrected in time	3	1	1	1	3	3
<i>Independence -technical</i>						
Different hardware/media	3	3	5	1	3	3
Different operating system	5	5	5	1	5	5
Different software	3	5	5	1	5	5
<i>Independence -organisation</i>						
Different internal damage preventions	3	3	3	1	3	3
Different war/terror attacks damage prev.	3	3	1	1	5	5
Different virus, worms attacks damage prev.	3	3	5	1	3	3
Different natural disaster damage prev.	3	3	3	1	5	5

Table 2 Uniform scale values for general SLA results used as input to Plato

On this basis the general results from Plato are calculated. The results are given in Table 3.

Requirement	SLA case					
	S1 DK	S2 Full replica -> checksum	S3 For confi- dential material	S4 Full replica on cloud	S5 For bit safe material	S6 As S5 with extra checksum
Confidentiality (totals)	1,50	1,50	2,50	0,50	1,25	1,25
Authorisation security violation	0,75	0,75	1,25	0,25	0,25	0,25
Physical security violation	0,75	0,75	1,25	0,25	0,75	0,75
Technical security violation	0,75	0,75	1,25	0,25	0,75	0,75
Transmission security violation	0,75	0,75	1,25	0,25	0,75	0,75
Integrity (totals)	1,58	1,42	1,50	0,75	1,79	2,04
Bit errors are found	1,50	1,50	1,50	1,50	1,50	2,5
Bit errors are corrected in time	1,50	0,50	0,50	0,50	1,50	1,50
Different hardware/media	0,99	0,99	1,65	0,33	0,99	0,99
Different operating system	1,70	1,70	1,70	0,34	1,70	1,70
Different software	0,99	1,65	1,65	0,33	1,65	1,65
Different internal damage preventions	0,75	0,75	0,75	0,25	0,75	0,75
Different war/terror attacks damage prev.	0,75	0,75	0,25	0,25	1,25	1,25
Different virus, worms attacks damage prev.	0,75	0,75	1,25	0,25	0,75	0,75
Different natural disaster damage prev.	0,75	0,75	0,75	0,25	1,25	1,25
Totals	3,08	2,92	4,00	1,25	3,04	3,29

Table 3 Plato results for SLA cases in general

The totals are the same as given in table 5 in Paper E.

1.3 Material specific results

When we pass input for specific material to Plato, we first have to define how we rate the different fulfillment of the requirements. We therefore define values in the uniform scale corresponding to how we rate values of probability for fulfillment compared to how important the requirement was rated to be. This is illustrated in Figure 30.

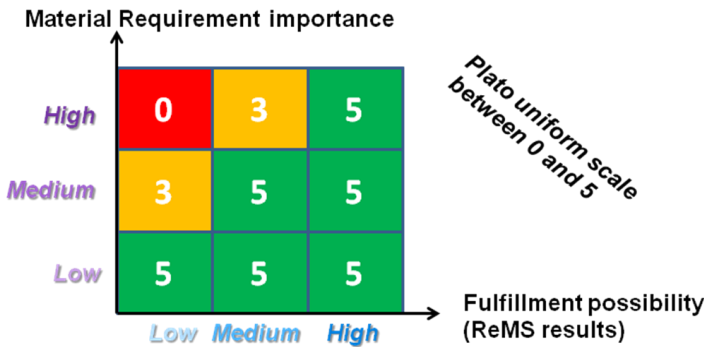


Figure 30 Transforming scheme to Plato uniform scale

Note that this means that we do not accept a low probability for fulfillment of a requirement in case the requirement was set to be of high importance. This corresponds to table 6 in Paper E.

1.3.1 Plato results for digitally born diaries

The requirements for digitally born diaries were that we require both high confidentiality and high bit safety. The requirements related to the SLAs are the same for all SLAs, thus the requirements are given in Table 4 (corresponding to M1 column of Table 1 in Paper E).

Requirement	SLA case					
	S1 DK	S2 Full replica -> checksum	S3 For confidential material	S4 Full replica on cloud	S5 For bit safe material	S6 As S5 with extra checksum
Confidentiality (totals)						
Authorisation security violation	High	High	High	High	High	High
Physical security violation	High	High	High	High	High	High
Technical security violation	High	High	High	High	High	High
Transmission security violation	High	High	High	High	High	High
Integrity (totals)						
Bit errors are found	High	High	High	High	High	High
Bit errors are corrected in time	High	High	High	High	High	High
Different hardware/media	High	High	High	High	High	High
Different operating system	High	High	High	High	High	High
Different software	High	High	High	High	High	High
Different internal damage preventions	High	High	High	High	High	High
Different war/terror attacks damage prev.	High	High	High	High	High	High
Different virus, worms attacks damage prev.	High	High	High	High	High	High
Different natural disaster damage prev.	High	High	High	High	High	High

Table 4 Requirements for digitally born diaries given for each SLA

Using the scale values from Figure 30 we get the values given in Table 5.

Requirement	SLA case					
	S1 DK	S2 Full replica -> checksum	S3 For confi- dential material	S4 Full replica on cloud	S5 For bit safe material	S6 As S5 with extra checksum
Confidentiality (totals)						
Authorisation security violation	H-M=3	H-M=3	H-H=5	H-L=0	H-L=0	H-L=0
Physical security violation	H-M=3	H-M=3	H-H=5	H-L=0	H-M=3	H-M=3
Technical security violation	H-M=3	H-M=3	H-H=5	H-L=0	H-M=3	H-M=3
Transmission security violation	H-M=3	H-M=3	H-H=5	H-L=0	H-M=3	H-M=3
Integrity (totals)						
Bit errors are found	H-M=3	H-M=3	H-M=3	H-M=3	H-M=3	H-H=5
Bit errors are corrected in time	H-M=3	H-L=0	H-L=0	H-L=0	H-M=3	H-M=3
Different hardware/media	H-M=3	H-M=3	H-H=5	H-L=0	H-M=3	H-M=3
Different operating system	H-H=5	H-H=5	H-H=5	H-L=0	H-H=5	H-H=5
Different software	H-M=3	H-H=5	H-H=5	H-L=0	H-H=5	H-H=5
Different internal damage preventions	H-M=3	H-M=3	H-M=3	H-L=0	H-M=3	H-M=3
Different war/terror attacks damage prev.	H-M=3	H-M=3	H-L=0	H-L=0	H-H=5	H-H=5
Different virus, worms attacks damage prev.	H-M=3	H-M=3	H-H=5	H-L=0	H-M=3	H-M=3
Different natural disaster damage prev.	H-M=3	H-M=3	H-M=3	H-L=0	H-H=5	H-H=5

Table 5 Uniform scale values for digitally born diaries used as input to Plato

Plato gives the results listed in Table 6

Requirement	SLA case					
	S1 DK	S2 Full replica -> checksum	S3 For confi- dential material	S4 Full replica on cloud	S5 For bit safe material	S6 As S5 with extra checksum
Confidentiality (totals)	1,50	-	-	-	-	-
Authorisation security violation	0,75	-	-	*	*	*
Physical security violation	0,75	-	-	*	-	-
Technical security violation	0,75	-	-	*	-	-
Transmission security violation	0,75	-	-	*	-	-
Integrity (totals)	1,58	-	-	-	-	-
Bit errors are found	1,50	-	-	-	-	-
Bit errors are corrected in time	1,50	*	*	*	-	-
Different hardware/media	0,99	-	-	*	-	-
Different operating system	1,70	-	-	*	-	-
Different software	0,99	-	-	*	-	-
Different internal damage preventions	0,75	-	-	*	-	-
Different war/terror attacks damage prev.	0,75	-	*	*	-	-
Different virus, worms attacks damage prev.	0,75	-	-	*	-	-
Different natural disaster damage prev.	0,75	-	-	*	-	-
Totals	3,08	-	-	-	-	-

Table 6 Plato results for SLA cases for digitally born diaries

The totals are the same as given in table 7 in Paper E. Note that the pattern is the same as for the general results, but the choice of using the uniform scale value 0 makes it discharge SLAs with value 0 for any of the requirements (marked with stars and red colour).

1.3.2 Plato results for digitally born images

The requirements for digitally born images were that we require high bit safety, but no confidentiality. The requirements related to the SLAs are the same for all SLAs, thus the requirements are given in Table 7 (corresponding to M2 column of Table 1 in Paper E).

Requirement	SLA case					
	S1 DK	S2 Full replica -> checksum	S3 For confi- dential material	S4 Full replica on cloud	S5 For bit safe material	S6 As S5 with extra checksum
Confidentiality (totals)						
Authorisation security violation	Low	Low	Low	Low	Low	Low
Physical security violation	Low	Low	Low	Low	Low	Low
Technical security violation	Low	Low	Low	Low	Low	Low
Transmission security violation	Low	Low	Low	Low	Low	Low
Integrity (totals)						
Bit errors are found	High	High	High	High	High	High
Bit errors are corrected in time	High	High	High	High	High	High
Different hardware/media	High	High	High	High	High	High
Different operating system	High	High	High	High	High	High
Different software	High	High	High	High	High	High
Different internal damage preventions	High	High	High	High	High	High
Different war/terror attacks damage prev.	High	High	High	High	High	High
Different virus, worms attacks damage prev.	High	High	High	High	High	High
Different natural disaster damage prev.	High	High	High	High	High	High

Table 7 Requirements for digitally born images given for each SLA

Using the scale values from Figure 30 we get the values given in Table 8.

Requirement	SLA case					
	S1 DK	S2 Full replica -> checksum	S3 For confi- dential material	S4 Full replica on cloud	S5 For bit safe material	S6 As S5 with extra checksum
Confidentiality (totals)						
Authorisation security violation	L-M=5	L-M=5	H-H=5	L-L=5	L-L=5	L-L=5
Physical security violation	L-M=5	L-M=5	H-H=5	L-L=5	L-M=5	L-M=5
Technical security violation	L-M=5	L-M=5	H-H=5	L-L=5	L-M=5	L-M=5
Transmission security violation	L-M=5	L-M=5	H-H=5	L-L=5	L-M=5	L-M=5
Integrity (totals)						
Bit errors are found	H-M=3	H-M=3	H-M=3	H-M=3	H-M=3	H-H=5
Bit errors are corrected in time	H-M=3	H-L=0	H-L=0	H-L=0	H-M=3	H-M=3
Different hardware/media	H-M=3	H-M=3	H-H=5	H-L=0	H-M=3	H-M=3
Different operating system	H-H=5	H-H=5	H-H=5	H-L=0	H-H=5	H-H=5
Different software	H-M=3	H-H=5	H-H=5	H-L=0	H-H=5	H-H=5
Different internal damage preventions	H-M=3	H-M=3	H-M=3	H-L=0	H-M=3	H-M=3
Different war/terror attacks damage prev.	H-M=3	H-M=3	H-L=0	H-L=0	H-H=5	H-H=5
Different virus, worms attacks damage prev.	H-M=3	H-M=3	H-H=5	H-L=0	H-M=3	H-M=3
Different natural disaster damage prev.	H-M=3	H-M=3	H-M=3	H-L=0	H-H=5	H-H=5

Table 8 Uniform scale values for digitally born images used as input to Plato

Plato gives the results listed in Table 9.

Requirement	SLA case					
	S1 DK	S2 Full replica -> checksum	S3 For confi- dential material	S4 Full replica on cloud	S5 For bit safe material	S6 As S5 with extra checksum
Confidentiality (totals)	2,50	-	-	-	2,50	2,50
Authorisation security violation	1,25	-	-	*	1,25	1,25
Physical security violation	1,25	-	-	*	1,25	1,25
Technical security violation	1,25	-	-	*	1,25	1,25
Transmission security violation	1,25	-	-	*	1,25	1,25
Integrity (totals)	1,58	-	-	-	1,79	2,04
Bit errors are found	1,50	-	-	-	1,50	2,5
Bit errors are corrected in time	1,50	*	*	*	1,50	1,50
Different hardware/media	0,99	-	-	*	0,99	0,99
Different operating system	1,70	-	-	*	1,70	1,70
Different software	0,99	-	-	*	1,65	1,65
Different internal damage preventions	0,75	-	-	*	0,75	0,75
Different war/terror attacks damage prev.	0,75	-	*	*	1,25	1,25
Different virus, worms attacks damage prev.	0,75	-	-	*	0,75	0,75
Different natural disaster damage prev.	0,75	-	-	*	1,25	1,25
Totals	4,08	-	-	-	4,29	4,54

Table 9 Plato results for SLA cases for digitally born images

The totals are the same as given in table 8 in Paper E.

I.3.3 Plato results for digitised books

The requirements for digitised books were that there were no requirements for confidentiality, and only some requirements to bit safety. The requirements are given in Table 10 (corresponding to M3 column of Table 1 in Paper E).

Requirement	SLA case					
	S1 DK	S2 Full replica -> checksum	S3 For confi- dential material	S4 Full replica on cloud	S5 For bit safe material	S6 As S5 with extra checksum
Confidentiality (totals)						
Authorisation security violation	Low	Low	Low	Low	Low	Low
Physical security violation	Low	Low	Low	Low	Low	Low
Technical security violation	Low	Low	Low	Low	Low	Low
Transmission security violation	Low	Low	Low	Low	Low	Low
Integrity (totals)						
Bit errors are found	Medium	Medium	Medium	Medium	Medium	Medium
Bit errors are corrected in time	High	High	High	High	High	High
Different hardware/media	Low	Low	Low	Low	Low	Low
Different operating system	Medium	Medium	Medium	Medium	Medium	Medium
Different software	Medium	Medium	Medium	Medium	Medium	Medium
Different internal damage preventions	Medium	Medium	Medium	Medium	Medium	Medium
Different war/terror attacks damage prev.	Low	Low	Low	Low	Low	Low
Different virus, worms attacks damage prev.	High	High	High	High	High	High
Different natural disaster damage prev.	Medium	Medium	Medium	Medium	Medium	Medium

Table 10 Requirements for digitised books given for each SLA

Using the scale values from Figure 30 we get the values given in Table 11.

Requirement	SLA case					
	S1 DK	S2 Full replica -> checksum	S3 For confi- dential material	S4 Full replica on cloud	S5 For bit safe material	S6 As S5 with extra checksum
Confidentiality (totals)						
Authorisation security violation	L-M=5	L-M=5	H-H=5	L-L=5	L-L=5	L-L=5
Physical security violation	L-M=5	L-M=5	H-H=5	L-L=5	L-M=5	L-M=5
Technical security violation	L-M=5	L-M=5	H-H=5	L-L=5	L-M=5	L-M=5
Transmission security violation	L-M=5	L-M=5	H-H=5	L-L=5	L-M=5	L-M=5
Integrity (totals)						
Bit errors are found	M-M=5	M-M=5	M-M=5	M-M=5	M-M=5	M-H=5
Bit errors are corrected in time	H-M=3	H-L=0	H-L=0	H-L=0	H-M=3	H-M=3
Different hardware/media	L-M=5	L-M=5	L-H=5	L-L=5	L-M=5	L-M=5
Different operating system	M-H=5	M-H=5	M-H=5	M-L=3	M-H=5	M-H=5
Different software	M-M=5	M-H=5	M-H=5	M-L=3	M-H=5	M-H=5
Different internal damage preventions	M-M=5	M-M=5	M-M=5	M-L=3	M-M=5	M-M=5
Different war/terror attacks damage prev.	L-M=5	L-M=5	L-L=5	L-L=5	L-H=5	L-H=5
Different virus, worms attacks damage prev.	H-M=3	H-M=3	H-H=5	H-L=0	H-M=3	H-M=3
Different natural disaster damage prev.	M-M=5	M-M=5	M-M=5	M-L=3	M-H=5	M-H=5

Table 11 Uniform scale values for digitised books used as input to Plato

Plato gives the results listed in Table 12

Requirement	SLA case					
	S1 DK	S2 Full replica -> checksum	S3 For confi- dential material	S4 Full replica on cloud	S5 For bit safe material	S6 As S5 with extra checksum
Confidentiality (totals)	2,50	-	-	-	2,50	2,50
Authorisation security violation	1,25	-	-	-	1,25	1,25
Physical security violation	1,25	-	-	-	1,25	1,25
Technical security violation	1,25	-	-	-	1,25	1,25
Transmission security violation	1,25	-	-	-	1,25	1,25
Integrity (totals)	2,19	-	-	-	2,19	2,19
Bit errors are found	2,50	-	-	-	2,50	2,5
Bit errors are corrected in time	1,50	*	*	*	1,50	1,50
Different hardware/media	1,65	-	-	-	1,65	1,65
Different operating system	1,70	-	-	-	1,70	1,70
Different software	1,65	-	-	-	1,65	1,65
Different internal damage preventions	1,25	-	-	-	1,25	1,25
Different war/terror attacks damage prev.	1,25	-	-	-	1,25	1,25
Different virus, worms attacks damage prev.	0,75	-	-	*	0,75	0,75
Different natural disaster damage prev.	1,25	-	-	-	1,25	1,25
Totals	4,69	-	-	-	4,69	4,69

Table 12 Plato results for SLA cases for digitised books

The totals are the same as given in table 9 in Paper E.

Appendix II. The BR-ReMS User interface

This appendix contains screen dumps of the BR-ReMS User interface represented in forms. Each form is followed by a description of the fields and command buttons in the form. Throughout the appendix there will be references to tables shown in Appendix III “The BR-ReMS Data Model”.

Table of contents of the appendix:

Main form.....	174
II.1. Forms for User Requirements Specifications	175
II.1.1 User requirements	176
II.1.2 Intermediate results in result characteristics	177
<i>System result characteristics</i>	178
<i>Functions for system result characteristics</i>	179
<i>Pillar result characteristics</i>	180
<i>Function for pillar result characteristics</i>	181
II.2. Forms for Service Level Agreement Specifications	183
II.2.1 Service level agreements	183
II.2.2 Service level agreements systems characteristics	184
<i>Service level agreements systems characteristics</i>	185
<i>Values of service level agreements systems characteristics</i>	185
<i>Service level agreements pillar characteristics</i>	186
<i>Values of service level agreements pillar characteristics</i>	186
II.3. Forms for Bit Repository Specifications	187
II.3.1 BR System Characteristics	188
II.3.2 Values of BR System Characteristics	189
II.3.3 BR Pillars.....	189
II.3.4 BR Pillar Characteristics.....	190
II.3.5 Values of BR Pillar Characteristics.....	191
II.4. Forms for Calculations.....	192
II.4.1 Values of intermediate results on pillar level	193
<i>Values of intermediate results for a specific pillar</i>	194
II.4.2 Values of intermediate results for accumulated values on system level.....	195
II.4.3 Values of intermediate results on system level	196
II.4.4 Final results for user requirements.....	197
II.5. Miscellaneous.....	198
II.5.1 Value type descriptions.....	198
II.5.2 Overview of all pillar values for one characteristic.....	198
II.5.3 Overview of all systems characteristics	199
II.5.4 Overview of all pillar characteristics	200
II.5.5 Queries	200
II.5.6 Reports	200

The main form, which is the starting point, is illustrated in Figure 31.

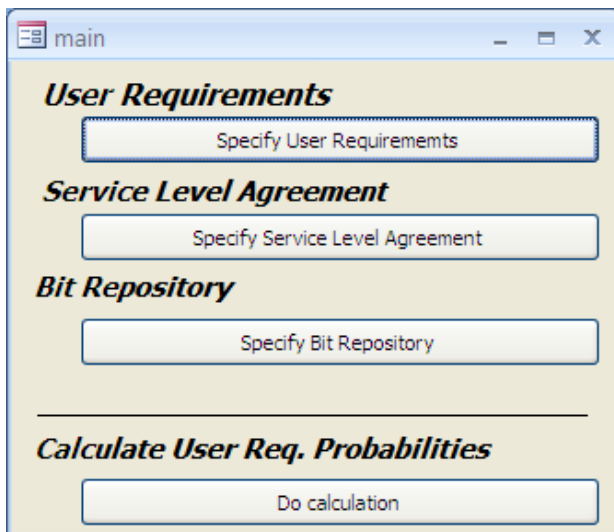


Figure 31 The BR-ReMS main form

Click on command button *Specify User Requirements* will open the form described in section II.1.

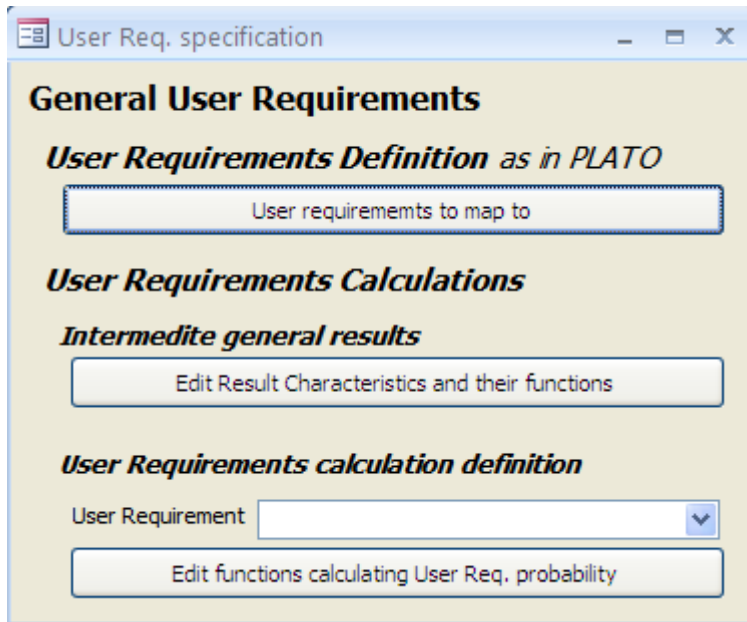
Click on command button *Specify Service Level Agreement* will open the form described in section II.2.

Click on command button *Specify Bit Repository* will open the form described in section II.3.

Click on command button *Do Calculations* will open the form described in section II.4.

II.1 Forms for User Requirements Specifications

The 'The BR-ReMS main form for requirements specification' is given in Figure 32. The form is activated by click on *Specify Bit Repository* in the main form shown in Figure 31.



The screenshot shows a window titled "User Req. specification". The main content area has a light yellow background and is titled "General User Requirements". It is divided into three sections:

- User Requirements Definition as in PLATO**: Contains a text input field with the placeholder text "User requirements to map to".
- User Requirements Calculations**: Contains a button labeled "Intermedite general results". Below this is a sub-section titled "User Requirements calculation definition" which includes a dropdown menu labeled "User Requirement" and a button labeled "Edit functions calculating User Req. probability".

Figure 32 The BR-ReMS main form for requirements specification

Click on command button *User Requirements to map to* will open the form described in section II.1.1

Click on command button *Edit Result Characteristics and their functions* will open the form described in section II.1.2.

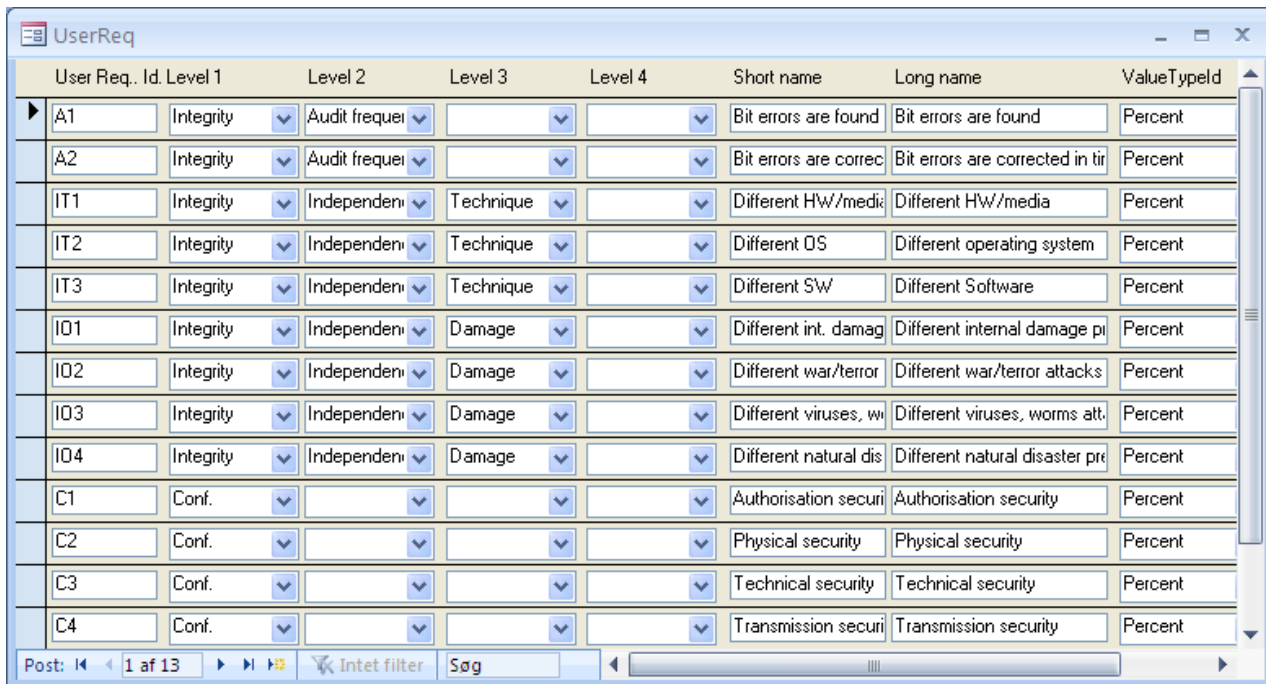
The drop down field *User Requirement* gives option to choose between *User Requirements to map to*.

These values are based on values from the *UserReq* table.

Click on command button *Edit functions calculating User Req. probability* will open the form described in section II.1.3 giving the User Requirement as parameter.

II.1.1 User requirements

Click on *User requirements to map to* on the form from Figure 32, will open the form 'The requirements form for definition of requirements' shown in Figure 33.



User Req. Id.	Level 1	Level 2	Level 3	Level 4	Short name	Long name	ValueTypeId
A1	Integrity	Audit frequen			Bit errors are found	Bit errors are found	Percent
A2	Integrity	Audit frequen			Bit errors are correc	Bit errors are corrected in tir	Percent
IT1	Integrity	Independen	Technique		Different HW/medi	Different HW/media	Percent
IT2	Integrity	Independen	Technique		Different OS	Different operating system	Percent
IT3	Integrity	Independen	Technique		Different SW	Different Software	Percent
I01	Integrity	Independen	Damage		Different int. damag	Different internal damage pi	Percent
I02	Integrity	Independen	Damage		Different war/terror	Different war/terror attacks	Percent
I03	Integrity	Independen	Damage		Different viruses, w	Different viruses, worms att.	Percent
I04	Integrity	Independen	Damage		Different natural dis	Different natural disaster pri	Percent
C1	Conf.				Authorisation securi	Authorisation security	Percent
C2	Conf.				Physical security	Physical security	Percent
C3	Conf.				Technical security	Technical security	Percent
C4	Conf.				Transmission securi	Transmission security	Percent

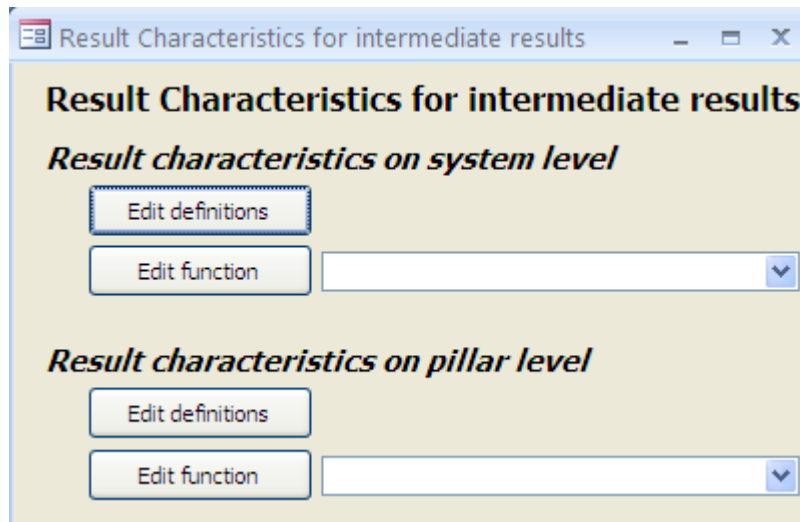
Figure 33 The requirements form for definition of requirements

All fields are changeable.

These values are represents entries in the *UserReq* table, where *Level1*, *Level2*, *Level3* and *Level4* are looked up in the *UserReqLevelName* table, and the *ValueTypeId* is looked up in the *ValueType* table.

II.1.2 Intermediate results in result characteristics

Click on *Edit result characteristics and their functions* on the form from Figure 32, will open the form 'Requirements main form for specification of result characteristics' shown in Figure 34.



The screenshot shows a software window titled "Result Characteristics for intermediate results". The window contains two main sections. The first section is titled "Result characteristics on system level" and contains two buttons: "Edit definitions" and "Edit function" next to a dropdown menu. The second section is titled "Result characteristics on pillar level" and also contains two buttons: "Edit definitions" and "Edit function" next to a dropdown menu. The form has a light beige background and a blue border.

Figure 34 Requirements main form for specification of result characteristics

Click on command button *Edit definitions* under title 'Result characteristics on **system level**' will open the form described in section "System result characteristics" given below.

Click on command button *Edit function* under title 'Result characteristics on **system level**' will open the form described in section "Functions for system result characteristics" given below. The value of the related drop down box, which gives choices of result characteristics on system level, is given as parameter to the form. The values in the drop down box are based on the *Char* table with selection of characteristics with type "FCTSY" which is the type for result characteristics on system level.

Click on command button *Edit definitions* under title 'Result characteristics on **pillar level**' will open the form described in section "Pillar result characteristics" given below.

Click on command button *Edit function* under title 'Result characteristics on **pillar level**' will open the form described in section "Functions for pillar result characteristics" given below. The value of the related drop down box, which gives choices of result characteristics on pillar level, is given as parameter to the form. The values in the drop down box are based on the *Char* table with selection of characteristics with type "FCTPL" which is the type for result characteristics on pillar level.

System result characteristics

Click on *Edit definitions* under *system level* on the form from Figure 34, will open the form 'The form for definition of system level result characteristics' shown in Figure 35.

Pillar Char. Id.	Level 1	Level 2	Level 3	Short name	Long name	ValueTypeId
FCT1L	Technique			Bit audit time	Bit audit time	No hours for c
FCT10L	Organisation			Average military disl	Average distance to military	No. of KM
FCT20L	Organisation			Average political dis	Average distance to politica	No. of KM
FCT11L	Organisation			Min military distance	Minimum distance to military	No. of KM
FCT12L	Organisation			Max military distanc	Maximum distance to militar	No. of KM
FCT21L	Organisation			Min political distanc	Minimum distance to politic	No. of KM
FCT22L	Organisation			Max political distanc	Maximum distance to politic	No. of KM
FCT40L	Organisation			Average distance b	Average distance between	No. of KM
FCT41L	Organisation			Min distance betwe	Minimum distance between	No. of KM
FCT42L	Organisation			Max distance betwe	Maximum distance between	No. of KM
*						

Figure 35 The form for definition of system level result characteristics

All fields are changeable.

These values represent entries in the *Char* table, where *Level1*, *Level2* and *Level3* are looked up in the *CharLevelName* table, and the *ValueTypeId* is looked up in the *ValueType* table. The only entries shown are entries with char type "FCTSY" which is the type for result characteristics on system level.

Functions for system result characteristics

Choice of a characteristic and click on *Edit functions* under *system level* on the form from Figure 34, will open the form 'The form for specification of function to system level result characteristics' shown in Figure 36.

Accumulated pillar result in separate result characteristic

Function for result characteristic on System level

Possible distruction by political attack

Involved system Characteristica

Min political distance
Average political distance

Add System Char
Remove Sys Char

Involved pillar Characteristica

No of full repl

Add Pillar Char
Remove Pillar Char

Fuction to calculate 'Possible distruct by political att' based on selected charac

```
IF <No of full repl> = 1 THEN ThisValue:=0,95
ELSEIF <Avarage political distance> < 5 THEN ThisValue:=0,95
ELSEIF <Avarage political distance> < 30 THEN ThisValue:=0,75
ELSEIF <Min political distance> > 150 THEN ThisValue:=0,75
ELSEIF [<Avarage political distance> > 1000] AND <No of full repl> > 3 THEN ThisValue:=0,05
ELSEIF [<Avarage political distance> > 150] THEN ThisValue:=0,25
ELSE ThisValue:=0,50
```

Figure 36 The form for specification of function to system level result characteristics

The data shown are based on data from the *FctResCharSysFunction* table, with arguments (characteristics) contained in the *FctResCharSubChar* table for the *CharId* given to the form.

The top drop down list shows information about the given *CharId* based on information from the *Char* table.

The list of *Involved system characteristics* contains the chosen system characteristics. These values are based on *FctResCharSubChar* table with *CharId* as the given *CharId* and *SubCharId* as the shown *CharId* in the related list (based on information from the *Char* table).

Click on the *Add System Char* command button results in adding the characteristic chosen in the related drop down box to the list. The only entries shown in the related drop down box are entries from the *Char* table with *CharType* "FCTSY", "SASY" or "BRSY" which are the types for system level char's.

Click on the *Remove Sys Char* command button results in removing the highlighting system char in the list.

The list of *Involved pillar characteristics* contains the chosen pillar characteristics. These values are based on *FctResCharSubChar* table with *CharId* as the given *CharId* and *SubCharId* as the shown *CharId* in the related list (based on information from the *Char* table).

Click on the *Add Pillar Char* command button results in adding the characteristic chosen in the related drop down box to the list. The only entries shown in the related drop down box are entries from the *Char* table with *CharType* "FCTPL", "SAPL" or "BRPL" which are the types for pillar level char's.

Click on the *Remove Pillar Char* command button results in removing the highlighting pillar char in the list.

The text field *Function to calculate 'Possible distruct by political att' based on selected char* is editable. It is based on the *Function* field in the *FctResCharSysFunction* table.

Pillar result characteristics

Click on *Edit definitions* under *pillar level* on the form from Figure 34, will open the form ‘The form for function to pillar level result characteristics’ shown in Figure 37.

Pillar Char. Id.	Level 1	Level 2	Level 3	Short name	Long name	ValueTypeId
FCT30	Technique			No of full repl	Number of full replicas	Number as po
FCT3	Technique			Exists off-site copy	Exists off-site copy	Yes/No
*						

Figure 37 The form for specification of pillar level result characteristics

All fields are changeable.

These values represent entries in the *Char* table, where *Level1*, *Level2* and *Level3* are looked up in the *CharLevelName* table, and the *ValueTypeId* is looked up in the *ValueType* table. The only entries shown are entries from the *Char* table with *CharType* “FCTPL” which is the type for result characteristics on pillar level.

Function for pillar result characteristics

Choice of a characteristic and click on *Edit functions* under *system level* on the form from Figure 34, will open the form 'The form for function to pillar level result characteristics' shown in Figure 38.

Res. Char Acc. function specification

Function for result characteristic on Pillar level

Number of full replicas [dropdown] Type **Number as p** [dropdown]

Involved pillar Characteristica

Object type SAPL [dropdown] Add Pillar Char [button] [dropdown]
Remove Pillar Char [button]

Single pillar
IF <object type> = 'Checksum' THEN ThisValue := 0 ELSE ThisValue := 1

Accumulating pillars
SUM OF ThisValue FROM EACH PILLAR

Figure 38 The form for function to pillar level result characteristics

The data shown are based on data from the *FctPillarCharAccFunction* table, with arguments (characteristics) contained in the *FctResCharSubChar* table for the same *CharId* given to the form.

The top drop down list shows information about the given *CharId* based on information from the *Char* table.

The Type shows the *CharType* for the given *CharId* with type description from the *ValueType* table.

The list of *Involved pillar characteristics* contains the chosen pillar characteristics. These values are based on *FctResCharSubChar* table with *CharId* as the given *CharId* and *SubCharId* as the shown *CharId* in the list (based on information from the *Char* table). The example given in the list is the short name of the characteristic and the type of the characteristic.

Click on the *Add Pillar Char* command button results in adding the characteristic chosen in the related drop down box to the list. The only entries shown in the related drop down box are entries from the *Char* table with *CharType* "FCTPL", "SAPL" or "BRPL" which are the types for pillar level characteristics.

Click on the *Remove Pillar Char* command button results in removing the highlighting pillar characteristic in the list.

The text field *Single pillar* is editable. It is based on the *FunctionSinglePillar* field in the *FctPillarCharAccFunction* table.

The text field *Accumulated pillars* is editable. It is based on the *FunctionAddingPillar* field in the *FctPillarCharAccFunction* table.

II.1.3 User Requirements calculations

Choice of a user requirements and Click on *Edit functions calculating User Req. probability* in Figure 32, results in opening the form 'The requirements form for specification of calculations of user requirements' shown in Figure 39.

The screenshot shows a web application window titled "User Requirement Functions". The main heading is "Calculation for user requirement" with a dropdown menu set to "Different war/terror attacks preventions". Below this are three sections:

- System Level Characteristica:** A list of characteristics: "Possible distract by political att", "Possible distract by one att", and "Possible distract by military att". To the right are "Add Syst. Char" and "Remove Syst. Char" buttons, each with a dropdown menu.
- Pillar Level Characteristica:** An empty list box. To the right are "Add Pillar Char" and "Remove Pillar Char" buttons, each with a dropdown menu.
- General function involving above characteristics:** A text area containing the following logic function:

```
possibilityOfDestruction :=
  MAX OF (<Possible distract by political att>), <Possible distract by one att>), <Possible distract by political att>);
IF possibilityOfDestruction > 0,75 THEN
  thisValue := LOW
ELSEIF possibilityOfDestruction > 0,22 THEN
  thisValue := MEDIUM
ELSE
  thisValue := HIGH
```

Figure 39 The requirements form for specification of calculations of user requirements

The data shown are based on data from the *FctUserReq* table, with arguments (characteristics) contained in the *FctUserReqXChar* table for the *UserReqId* given to the form.

The top drop down list shows information about the given *UserReqId* based on information from the *UserReq* table.

The list of *System Level characteristics* contains the chosen pillar characteristics. These values are based on *FctUserReqXChar* table with *UserReqId* as the given *UserReqId* and *CharId* as the shown *CharId* in the list (based on information from the *Char* table).

Click on the *Add Syst. Char* command button results in adding the characteristic chosen in the related drop down box to the list. The only entries shown in the related drop down box are entries from the *Char* table with *CharType* "FCTSY", "SASY" or "BRSY" which are the types for system level char's.

Click on the *Remove Syst. Char* command button results in removing the highlighting system characteristic in the related list.

The list of *Pillar Level Characteristics* contains the chosen pillar characteristics. These values are based on *FctUserReqXChar* table with *UserReqId* as the given *UserReqId* and *CharId* as the shown *CharId* in the list (based on information from the *Char* table).

Click on the *Add Pillar Char* command button results in adding the characteristic chosen in the related drop down box to the list. The only entries shown in the related drop down box are entries from the *Char* table with *CharType* "FCTPL", "SAPL" or "BRPL" which are the types for pillar level char's.

Click on the *Remove Pillar Char* command button results in removing the highlighting pillar characteristic in the related list.

The text field *General function involving above characteristics* is editable. It is based on the *UserReqFunction* field in the *FctUserReq* table.

II.2 Forms for Service Level Agreement Specifications

The 'The BR-ReMS main form for SLA specification' is given in Figure 40. The form is activated by click on *Specify Service Level Agreements* in the main form shown in Figure 31.

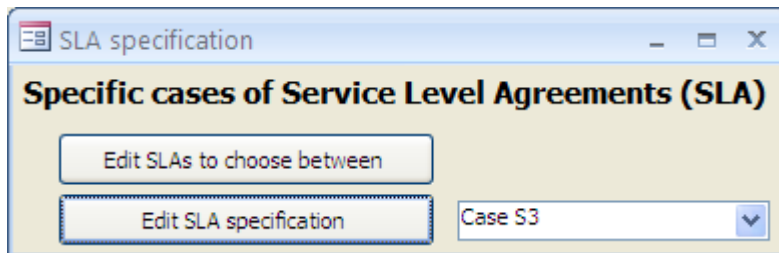


Figure 40 The BR-ReMS main form for SLA specification

Click on command button *Edit SLAs to choose between* will open the form described in section II.2.1
Click on command button *Edit SLA specification* will open the form described in section II.2.2, given the value of chosen SLA in the related drop down box. The values in the drop down box are based on entries in the *SLA* table.

II.2.1 Service level agreements

Click on *Edit SLAs to choose between* on the form from Figure 40, will open the form 'The SLA form for definition of SLAs' shown in Figure 41.

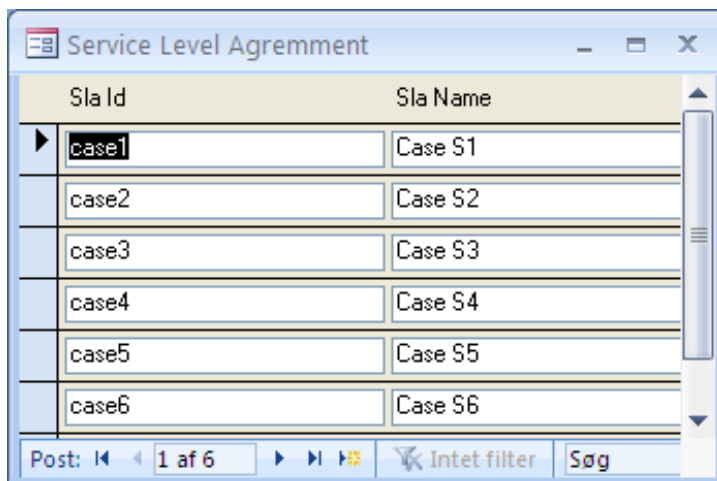


Figure 41 The SLA form for definition of SLAs

All fields are editable.

These values represent entries in the *SLA* table,

II.2.2 Service level agreements systems characteristics

Choice of SLA and click on *Edit SLAs specification* on the form from Figure 40, will open the form 'The SLA form for specification of a SLA' shown in Figure 42.

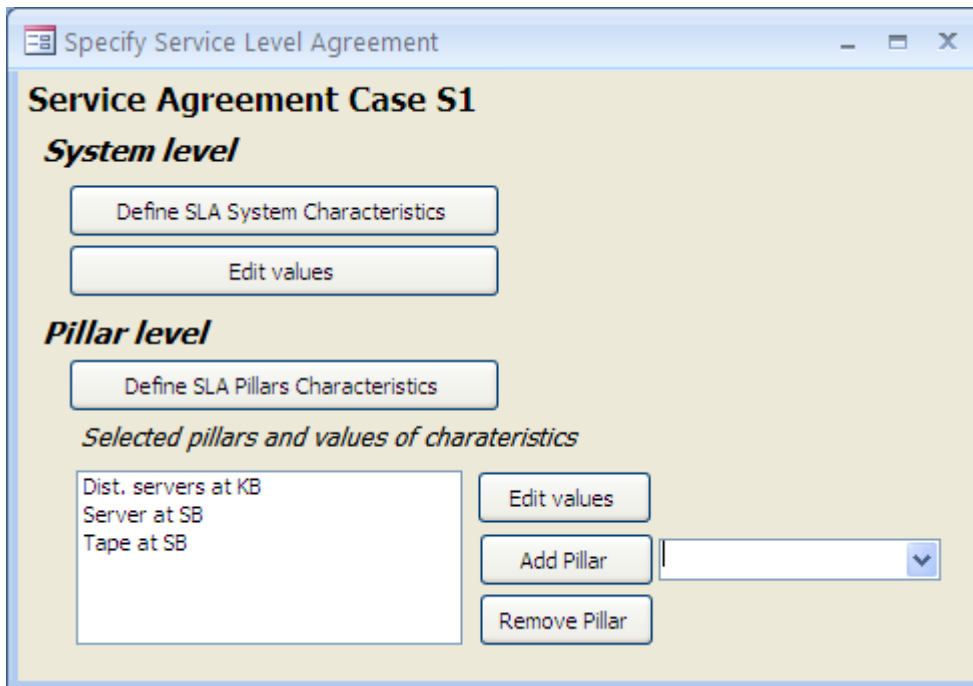


Figure 42 The SLA form for specification of a SLA

Click on command button *Define SLA system characteristics* will open the form described in section “Service level agreements system characteristics” given below.

Click on command button *Edit values* will open the form described in section “Values of service level agreements systems characteristics” given below

Click on command button *Define SLA pillar characteristics* will open the form described in section “Service level agreements pillar characteristics” given below.

The list of *Selected pillars and values of characteristics* contains the chosen pillars for the SLA. These values are based on *SlaPillar* table with *SlaId* as the given *SlaId* and *PillarId* as the shown *PillarId* in the list (based on information from the *BrPillar* table).

Click on the *Add Pillar* command button results in adding the pillar chosen in the related drop down box to the list. The entries shown in the related drop down box are entries from the *BrPillar* table.

Click on the *Remove Pillar* command button results in removing the highlighted pillar in the list.

Click on command button *Edit values* related to a highlighted pillar in the list will open the form described in section “Values of service level agreements pillar characteristics” given below. The value of chosen pillar in the pillar list is given as parameter to the form.

Service level agreements systems characteristics

Click on *Define SLA System Characteristics* on the form from Figure 42, will open the form 'The SLA specification form for specification of SLA system characteristics' shown in Figure 43.

Pillar Char. Id.	Level 1	Level 2	Level 3	Short name	Long name	ValueTypeId	Inc
SASYS3	Technique	Hardware		Tera bytes stored	Tera bytes stored	Number as po	Values
SASYS2	Technique	Application		Bit audit frequency	Bit audit frequency	No. of months	Values

Figure 43 The SLA specification form for specification of SLA system characteristics

All text fields are editable and values of drop down boxes can be changed.

These values represent entries in the *Char* table, where *Level1*, *Level2* and *Level3* are looked up in the *CharLevelName* table, and the *ValueTypeId* is looked up in the *ValueType* table. The only entries shown are entries with char type "SASY" which is the type for SLA characteristics on system level.

Click on one of the *Values* command buttons will open the form given in Figure 59 "Value type form with description of possible values for a type" shown in section II.5.1. This form gives a description of the possible values of the related *ValueTypeId*.

Values of service level agreements systems characteristics

Click on *Edit values* under *system level* on the form from Figure 42, will open the form 'The SLA specification form for values of SLA system characteristics' shown in Figure 44.

Group	Characteristic	Value	Type
Technique - Hardware	Tera bytes stored	10	Number a
Technique - Application	Bit audit frequency	4	No. of mor

Figure 44 The SLA specification form for values of SLA system characteristics

The *Value* fields are the only editable fields. The data is based on the *SlaSysCharVal* table.

Each line represents a characteristic with characteristic information based on the *Char* table, where *Group* is based on *L1Id*, *L2Id* and *L3Id* looked up in the *CharLevelName* table, and the *Type* is looked up in the *ValueType* table.

Click on one of the *Possible values* command buttons will open the form given in Figure 59 'Value type form with description of possible values for a type' shown in section II.5.1. This form gives the possible values for the related value type.

Service level agreements pillar characteristics

Click on *Define SLA Pillars Characteristics* on the form from Figure 42, will open the form 'The SLA specification form for specification of SLA pillar characteristics' shown in Figure 45.

Pillar Char. Id.	Level 1	Level 2	Level 3	Short name	Long name	ValueTypeId
TD01	Data	Type		Object type	Object type: Checksum or c	Object type Values
TD02	Data	Type		Checksum type	Checksum type	Bit check type Values

Figure 45 The SLA specification form for specification of SLA pillar characteristics

All text fields are editable and values of drop down boxes can be changed.

These values represent entries in the *Char* table, where *Level1*, *Level2* and *Level3* are looked up in the *CharLevelName* table, and the *ValueTypeId* is looked up in the *ValueType* table. The only entries shown are entries with char type "SAPL" which is the type for SLA characteristics on pillar level.

Click on one of the *Values* command buttons will open the form given in Figure 59 'Value type form with description of possible values for a type' shown in section II.5.1. This form gives a description of the possible values of the related *ValueTypeId*.

Values of service level agreements pillar characteristics

Choice of a pillar and click on *Edit values* under *pillar level* on the form from Figure 42, will open the form 'The SLA specification form for values of SLA pillar characteristics' shown in Figure 46.

Group	Characteristic	Value	Type
Data - Type	Object type: Checksum or copy of data	Checksum	Object typ
Data - Type	Checksum type	MD5	Bit check

Figure 46 The SLA specification form for values of SLA pillar characteristics

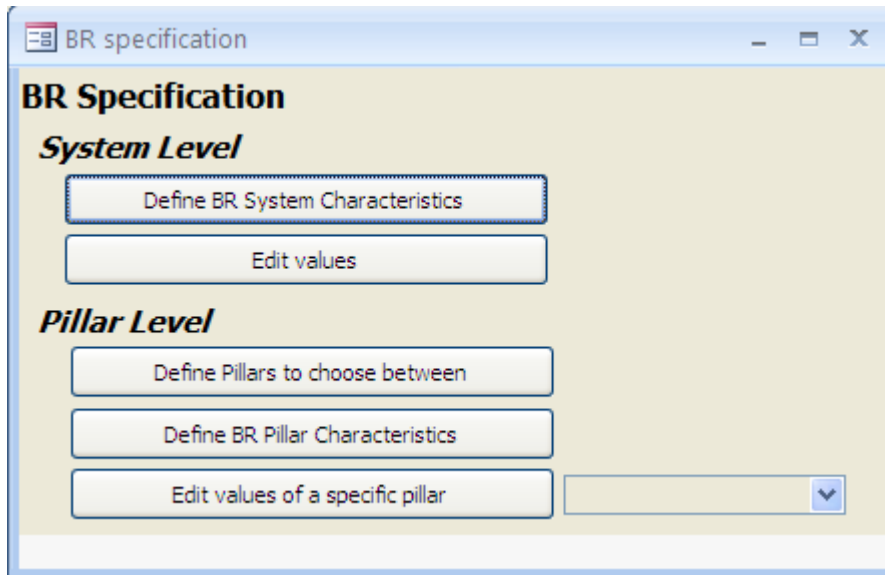
The *Value* fields are the only editable fields. The data is based on the *SlaPillarCharVal* table.

Each line represents a characteristic with characteristic information based on the *Char* table, where *Group* is based on *L1Id*, *L2Id* and *L3Id* looked up in the *CharLevelName* table, and the *Type* is looked up in the *ValueType* table.

Click on one of the *Possible values* command buttons will open the form given in Figure 59 'Value type form with description of possible values for a type' shown in section II.5.1. This form gives a description of the possible values for the related value.

II.3 Forms for Bit Repository Specifications

The 'The BR-ReMS main form for the BR implementation' is given in Figure 47. The form is activated by click on *Specify Bit Repository* in the main form shown in Figure 31.



The screenshot shows a window titled "BR specification" with a light yellow background. The window is divided into two main sections: "System Level" and "Pillar Level".

- System Level:** Contains two buttons: "Define BR System Characteristics" (with a dotted border) and "Edit values".
- Pillar Level:** Contains three buttons: "Define Pillars to choose between", "Define BR Pillar Characteristics", and "Edit values of a specific pillar". To the right of the last button is a dropdown menu.

Figure 47 The BR-ReMS main form for the BR implementation

Click on command button *Define BR system characteristics* will open the form described in section II.3.1.

Click on command button *Edit values* will open the form described in section II.3.2

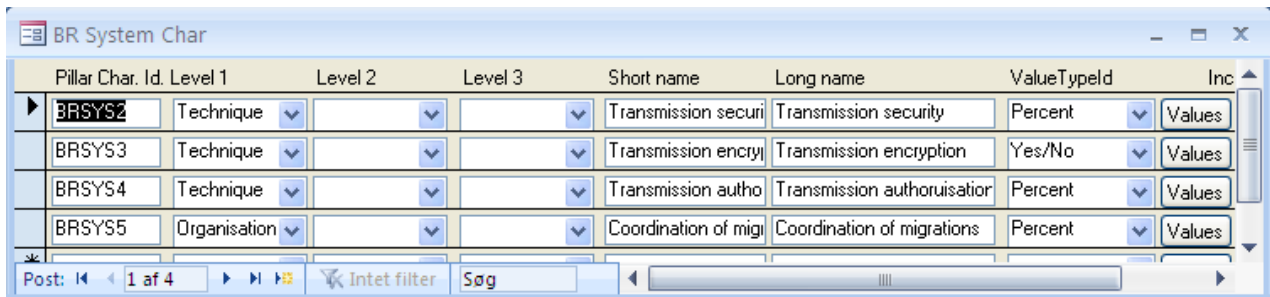
Click on command button *Define pillars to choose between* will open the form described in section II.3.3

Click on command button *Define BR pillar characteristics* will open the form described in section II.3.4

Click on command button *Edit values of a specific pillar* will open the form described in section II.3.5, given the value of chosen pillar in the related drop down box. The values in the drop down box are based on entries in the *BRPillar* table.

II.3.1 BR System Characteristics

Click on *Define BR System Characteristics* on the form from Figure 47, will open the form ‘The BR-ReMS BR form for specification of BR system characteristics’ shown in Figure 48.



The screenshot shows a web application window titled "BR System Char". It contains a table with the following columns: Pillar Char. Id., Level 1, Level 2, Level 3, Short name, Long name, ValueTypeId, and Inc. The table lists four entries:

Pillar Char. Id.	Level 1	Level 2	Level 3	Short name	Long name	ValueTypeId	Inc
BRSYS2	Technique			Transmission securi	Transmission security	Percent	Values
BRSYS3	Technique			Transmission encry	Transmission encryption	Yes/No	Values
BRSYS4	Technique			Transmission autho	Transmission authorisation	Percent	Values
BRSYS5	Organisation			Coordination of mig	Coordination of migrations	Percent	Values

Below the table, there is a navigation bar with "Post: 1 af 4", a search box labeled "Søg", and a "Intet filter" button.

Figure 48 The BR-ReMS BR form for specification of BR system characteristics

All text fields are editable and values of drop down boxes can be changed.

These values represent entries in the *Char* table, where *Level1*, *Level2* and *Level3* are looked up in the *CharLevelName* table, and the *ValueTypeId* is looked up in the *ValueType* table. The only entries shown are entries with char type “BRSY” which is the type for BR characteristics on system level.

Click on one of the Values command buttons will open the form given in Figure 59 ‘Value type form with description of possible values for a type’ shown in section II.5.1. This form gives a description of the possible values of the related *ValueTypeId*.

II.3.2 Values of BR System Characteristics

Click on *Edit Values* on the form from in Figure 47, will open the form ‘The BR form for editing values of BR system characteristics’ shown in Figure 49.

Group	Characteristic	Value	Possible values	Type
Technique	Transmission security	90	Possible values	Percent
Technique	Transmission encryption	Y	Possible values	Yes/No
Technique	Transmission authorisation	80	Possible values	Percent
Organisation	Coordination of migrations	80	Possible values	Percent

Figure 49 The BR form for editing values of BR system characteristics

The *Value* fields are the only editable fields. The data is based on the *BrSysCharVal* table.

Each line represents a characteristic with characteristic information based on the *Char* table, where *Group* is based on *L1Id*, *L2Id* and *L3Id* looked up in the *CharLevelName* table, and the *Type* is looked up in the *ValueType* table.

Click on one of the *Possible values* command buttons will open the form given in Figure 59 ‘Value type form with description of possible values for a type’ shown in section II.5.1. This form gives a description of the possible values for the related value

II.3.3 BR Pillars

Click on *Define Pillars to choose between* on the form from Figure 47, will open the form ‘The BR form for specification of BR pillars’ shown in Figure 50.

Pillar Id	Pillar name
DVD	DVD at SA
DIST	Dist. servers at KB
SERVER	Server at SB
TAPE	Tape at SB
CLOUD	Cloud somewhere
PARTNER	Partner in Austria

Figure 50 The BR form for specification of BR pillars

All fields are editable.

These values represent entries in the *BRPillar* table.

II.3.4 BR Pillar Characteristics

Click on *Define BR Pillar Characteristics* on the form from Figure 47, will open the form ‘The BR form for specification of BR pillar characteristics’ shown in Figure 51.

Pillar Char. Id.	Level 1	Level 2	Level 3	Short name	Long name	ValueTypeId
TA03b	Technique	Hardware		Calculation speed	Calculation speed	No hours for c
TA22	Technique	Hardware		HW robustness	Hardware robustness	Low/Medium/
TA03a	Technique	Hardware		HW vendor	Hardware vendor	HW vendor
TA30	Technique	Hardware		Media Tech	Media Technology	Media technol
TA31	Technique	Hardware		Media Format	Media Format	Media format
TA60	Technique	Hardware		Media migration inte	Media migration interval	No. of months
TA15	Technique	Hardware		MeanTimeToFail	Mean Time To Failure	No. of months
TA16	Technique	Hardware		Life time	Media life time	No. of months
ISTEC3_1	Technique	Hardware	Service loss	Air-con or water sup	Air-con or water supply	Low/Medium/
ISTEC3_2	Technique	Hardware	Service loss	Power supply	Power supply	Low/Medium/
TA05	Technique	Software		OS Tech	Operating system technolog	Operating Sys
TA032	Technique	Software		OS vendor	Operating system vendor	Operating Sys
TA08	Technique	Software		Bit corr. (RAID)	Internal bit correction (RAIC	Yes/No
TA09	Technique	Software		Calc.on checksum	Calculation on checksum ty	Bit check type
TA40	Technique	Software		Protective virus/wo	Protective virus/worm meas	Percent
TA01	Technique	Application		Encryption	Encryption	Yes/No
TA02	Technique	Application		Compression	Compression	Yes/No
TA04	Technique	Application		SW Tech	Program Software technolo	SW technolog

Figure 51 The BR form for specification of BR pillar characteristics

All text fields are editable and values of drop down boxes can be changed.

These values represent entries in the *Char* table, where *Level1*, *Level2* and *Level3* are looked up in the *CharLevelName* table, and the *ValueTypeId* is looked up in the *ValueType* table. The only entries shown are entries with char type “BRPL” which is the type for BR characteristics on pillar level.

Click on one of the Values command buttons will open the form given in Figure 59 ‘Value type form with description of possible values for a type’ shown in section II.5.1. This form gives a description of the possible values of the related *ValueTypeId*.

II.3.5 Values of BR Pillar Characteristics

Choice of a pillar and click on *Edit values of a specific pillar* on the form from Figure 47, will open the form ‘The BR form for specification of BR pillar characteristics’ shown in Figure 52.

The screenshot shows a web application window titled 'Pillar Characteristic Values'. The main content area displays a table with the following data:

Type	Group	Characteristic	Value	Val. Type	
BRPL	Technique - Hardware	Calculation speed	15	No hours l	Possible values
BRPL	Technique - Hardware	Hardware robustness	Medium	Low/Medi	Possible values
BRPL	Technique - Hardware	Hardware vendor	HP	HW vendi	Possible values
BRPL	Technique - Hardware	Media Technology	Magnetic	Media tec	Possible values
BRPL	Technique - Hardware	Media Format	Disk	Media forr	Possible values
BRPL	Technique - Hardware	Media migration interval	60	No. of moi	Possible values
BRPL	Technique - Hardware	Mean Time To Failure	60	No. of moi	Possible values
BRPL	Technique - Hardware	Media life time	60	No. of moi	Possible values
BRPL	Technique - Hardware - Service loss	Air-con or water supply	Medium	Low/Medi	Possible values
BRPL	Technique - Hardware - Service loss	Power supply	Medium	Low/Medi	Possible values
BRPL	Technique - Software	Operating system technology	MS-Win	Operating	Possible values
BRPL	Technique - Software	Operating system vendor	Microsoft	Operating	Possible values
BRPL	Technique - Software	Internal bit correction (RAID more copies)	Y	Yes/No	Possible values
BRPL	Technique - Software	Calculation on checksum type	MD5	Bit check	Possible values
BRPL	Technique - Software	Protective virus/worm measures	80	Percent	Possible values
BRPL	Technique - Application	Program Software technology	Java int	SW techn	Possible values

At the bottom of the window, there is a navigation bar with the text 'Post: 1 af 82', a search box containing 'Søg', and a 'Filtreret' button.

Figure 52 The BR-ReMS main form for editing values of BR pillar characteristics

The *Value* fields are the only editable fields. The data is based on the *BrPillarCharVal* table.

Each line represents a characteristic with characteristic information based on the *Char* table, where *Group* is based on *L1Id*, *L2Id* and *L3Id* looked up in the *CharLevelName* table, and the *Type* is looked up in the *ValueType* table.

Click on one of the *Possible values* command buttons will open the form given in Figure 59 ‘Value type form with description of possible values for a type’ shown in section II.5.1. This form gives a description of the possible values for the related value

II.4 Forms for Calculations

The 'The BR-ReMS main form for calculations' is shown in Figure 53. The form is activated by click on *Do calculation* in the main form shown in Figure 31.

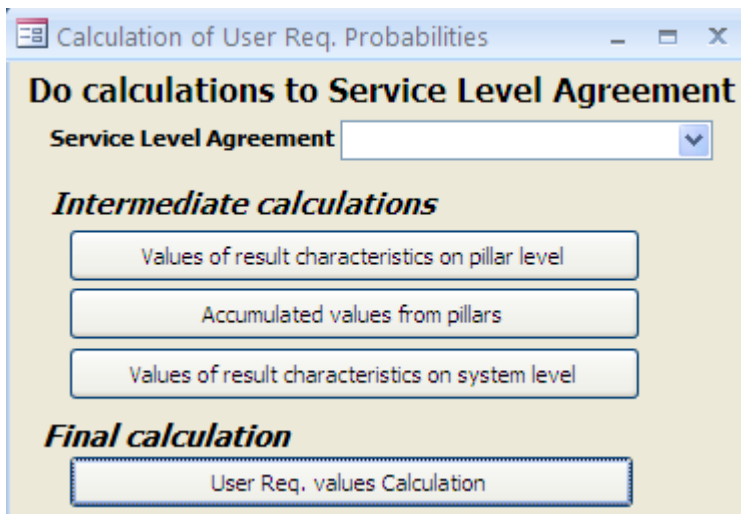


Figure 53 The BR-ReMS main form for calculations

The drop down box *Service Level Agreement* gives choice of the SLA that calculations concern. These values are based on values from the *SLa* table.

Click on command button *Values of result characteristics on pillar level* will open the form described in section II.4.1, given the value of chosen SLA in the drop down box.

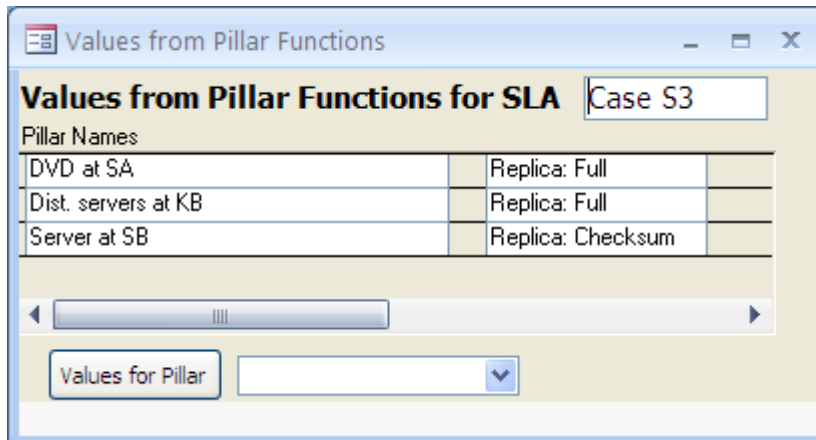
Click on command button *Accumulated values from pillars* will open the form described in section II.4.2, given the value of chosen SLA in the drop down box.

Click on command button *Values of result characteristics on system level* will open the form described in section II.4.3, given the value of chosen SLA in the drop down box.

Click on command button *User Req. values Calculation* will open the form described in section II.4.4, given the value of chosen SLA in the drop down box.

II.4.1 Values of intermediate results on pillar level

Choice of service level agreement and click on *Values of result characteristics on pillar level* on the form from Figure 53, will open the form 'The calculation form for intermediate results at pillar level' shown in Figure 54.



The screenshot shows a web application window titled "Values from Pillar Functions". The main heading is "Values from Pillar Functions for SLA" followed by a text input field containing "Case S3". Below this is a section labeled "Pillar Names" containing a table with three rows:

DVD at SA	Replica: Full
Dist. servers at KB	Replica: Full
Server at SB	Replica: Checksum

Below the table is a horizontal scroll bar. At the bottom left, there is a button labeled "Values for Pillar" and a dropdown menu.

Figure 54 The calculation form for intermediate results at pillar level

The top text field shows information about the given *Slaid* based on information from the *Slat* table.

The list *Pillar Names* shows the pillars chosen for the SLA and their replica type. These values are based on values from the *BRPillar* table joint with table *SlasysCharVal* table for the specific replica type *CharId* and for the given *Slaid*.

Click on command button *Values for pillar* will open the form described in section "Values of intermediate results for a specific pillar" given below. The value of chosen pillar in the related drop down box is given as parameter to the form. The values in the drop down list are based on values from the *BRPillar* table.

Values of intermediate results for a specific pillar

Choice of a pillar and click on *Values for pillar* on the form from Figure 54, will open the form ‘The calculation form for results for a specific pillar’ shown in Figure 55.

The screenshot shows a web application window titled "Values from Pillar Functions". The main heading is "Values for a Pillar in SLA" with a text input field containing "Case S3". Below this, there is a section for "Pillar Name" with a table listing "Dist. servers at KB" and "Replica: Full". A scrollable list follows. The next section is "Sla Pillar Char values" with a table containing two rows of data. The final section is "Pillar Char Values" with a table with columns "Group", "Characteristic", and "Value". Two rows are shown: "Technique" with "FCTPL" and "Number of full replicas" (value 1), and "Technique" with "FCTPL" and "Exists off-site copy" (value N). Each row has "Method" and "pillar values" buttons.

Figure 55 The calculation form for results for a specific pillar

The top text field shows information about the given *SlaId* based on information from the *Sla* table.

The list *Pillar Name* shows the pillars chosen for the SLA and their replica type. These values are based on values from the *BRPillar* table joint with table *SlaSysCharVal* table for the specific replica type *CharId* and for the given *SlaId*.

The list *Sla Pillar Char Values* shows the SLA pillar characteristics for the chosen SLA and pillar. This is shown since these can be parameters to the specified function (given via the *Method* buttons).

The *Pillar Char Value* is based on the *FctResPillarCharValue* table. The *Value* fields are the only editable fields. Each line represents a characteristic with characteristic information based on the *Char* table, where *Group* is based on *L1Id*, *L2Id* and *L3Id* looked up in the *CharLevelName* table, and the *Type* is looked up in the *ValueType* table.

Click on command buttons *Method* will open the form shown in Figure 38 ‘The form for function to pillar level result characteristics’ which gives the description of the function to calculate the value. The form is given the related characteristic.

Click on command buttons *pillar values* will open the form shown in Figure 52 ‘The BR-ReMS main form for editing values of BR pillar characteristics’. This option is there since all pillar characteristics can be parameters in the function specification via the *Method* buttons.

II.4.2 Values of intermediate results for accumulated values on system level

Choice of service level agreement and click on *Values of result characteristics on pillar level* on the form from Figure 53, will open the form 'The calculation form for intermediate accumulated results at system' shown in Figure 56.

Pillar Names		Replica
DVD at SA		Replica: Full
Dist. servers at KB		Replica: Full
Server at SB		Replica: Checksum

Pillar Char Values					
Group	Characteristic	Value	Method	pillar values	
Technique	FCTPL	Exists off-site copy	Y	Method	pillar values YN
Technique	FCTPL	Number of full replicas	2	Method	pillar values Number
Technique - Hardware	BRPL	Calculation speed	15-3000	Method	pillar values HoursTc
Technique - Hardware	BRPL	Hardware robustness	Medium+	Method	pillar values LMH
Technique - Hardware	BRPL	Hardware vendor	Varied	Method	pillar values HwVenc

Figure 56 The calculation form for intermediate accumulated results at system

The top text field shows information about the given *Slaid* based on information from the *Slat* table.

The list *Pillar Names* shows the pillars chosen for the SLA and their replica type. These values are based on values from the *BRPillar* table joint with table *SlasysCharVal* table for the specific replica type *CharId* and for the given *Slaid*.

The *Pillar Char Value* is based on the *FctPillarCharAccValue* table. The *Value* fields are the only editable fields. Each line represents a characteristic with characteristic information based on the *Char* table, where *Group* is based on *L1Id*, *L2Id* and *L3Id* looked up in the *CharLevelName* table, and the *Type* is looked up in the *ValueType* table.

Click on command buttons *Method* will open the form shown in Figure 38 'The form for function to pillar level result characteristics' which gives the description of the function to calculate the value. The form is given the related characteristic.

Click on command buttons *pillar values* will open the form shown in Figure 60 'Overview of all pillar values for one pillar characteristic and pillars in a SLA' in section II.5.2. This option is there since it is these values that are the parameters in the function specification via the *Method* buttons.

II.4.3 Values of intermediate results on system level

Choice of service level agreement and click on *Accumulated values from pillars* on the form from Figure 53, will open the form 'The calculation form for intermediate accumulated results at system' shown in Figure 57.

Group	Characteristic	Value	Method
Organisation	FCTSY Minimum distance to military t	20	NoKm
Organisation	FCTSY Possible distruction by one al	95	Percent
Organisation	FCTSY Minimum distance to political	0,5	NoKm
Organisation	FCTSY Maximum distance to politica	21	NoKm
Organisation	FCTSY Maximum distance between i	2	NoKm
Organisation	FCTSY Possible distruction by politic.	95	Percent

Figure 57 The calculation form for intermediate accumulated results at system

The top text field shows information about the given *Slaid* based on information from the *SLa* table.

The list *Pillar Names* shows the pillars chosen for the SLA. These values are based on values from the *BRPillar* table.

Click on command buttons *View pillar values* will open the form shown in Figure 52 'The BR-ReMS main form for editing values of BR pillar characteristics' for the related pillar. This option is there since all pillar characteristics can be parameters in the function specification via the *Method* buttons.

Click on command button *View General System Values* will open the form shown in Figure 62 'Overview of all pillar characteristics'. This option is there since all pillar characteristics can be parameters in the function specification via the *Method* buttons.

Click on command button *View Acc. Pillar Values* will open the form shown in Figure 61 'Overview of all systems characteristics' in section II.5.3. This option is there since all pillar characteristics can be parameters in the function specification via the *Method* buttons.

The *Pillar Char Value* is based on the *FctPillarCharAccValue* table. The *Value* fields are the only editable fields. Each line represents a characteristic with characteristic information based on the *Char* table, where *Group* is based on *L1Id*, *L2Id* and *L3Id* looked up in the *CharLevelName* table, and the *Type* is looked up in the *ValueType* table.

Click on command buttons *Method* will open the form shown in Figure 36 'The form for specification of function to system level result characteristics' which gives the description of the function to calculate the value. The form is given the related *CharId*.

II.4.4 Final results for user requirements

Choice of service level agreement and click on *User Req. values calculation* on the form from Figure 53, will open the form 'The calculation form for final user requirements results' shown in Figure 58.

The screenshot shows a web application window titled "User requirement values in Service Agreement". The main heading is "Service Level Agreement Case S3". Below this, there is a section for "Pillar Names" with a table:

Pillar Name	Replica Type
DVD at SA	Replica: Full
Dist. servers at KB	Replica: Full
Server at SB	Replica: Checksum

To the right of this table are two buttons: "View General System Values" and "View Acc. Pillar Values". Below the pillar names is a section for "User Req Values" with a table:

Level	User Req.	Value	Action
Integrity - Audit frequency -	Bit errors are found	Medium	View calculation instructions
Integrity - Audit frequency -	Bit errors are corrected in time	Low	View calculation instructions
Conf. - -	Authorisation security	High	View calculation instructions
Conf. - -	Physical security	High	View calculation instructions
Conf. - -	Technical security	High	View calculation instructions
Conf. - -	Transmission security	High	View calculation instructions
Integrity - Independence - Damage	Different internal damage preventions	Medium	View calculation instructions
Integrity - Independence - Damage	Different war/terror attacks preventio	Low	View calculation instructions
Integrity - Independence - Damage	Different viruses, worms attacks prev	High	View calculation instructions
Integrity - Independence - Damage	Different natural disaster preventions	Medium	View calculation instructions
Integrity - Independence - Technique	Different HW/media	High	View calculation instructions
Integrity - Independence - Technique	Different operating system	High	View calculation instructions
Integrity - Independence - Technique	Different Software	High	View calculation instructions

Figure 58 The calculation form for final user requirements results

The top text field shows information about the given *SlaId* based on information from the *Sla* table.

The list *Pillar Names* shows the pillars chosen for the SLA and their replica type. These values are based on values from the *BRPillar* table joint with table *SlaSysCharVal* table for the specific replica type *CharId* and for the given *SlaId*.

Click on command button *View General System Values* will open the form shown in Figure 62 'Overview of all pillar characteristics'. This option is there since all pillar characteristics can be parameters in the function specification via the *View calculation instructions* buttons.

Click on command button *View Acc. Pillar Values* will open the form shown in Figure 61 'Overview of all systems characteristics' in section II.5.3. This option is there since all pillar characteristics can be parameters in the function specification via the *View calculation instructions* buttons.

The *User Req Values* is based on the *UserReqValuesInSla* table. The *Value* fields are the only editable fields. Each line represents a user requirement with requirement information based on the *UserReq* table, where *Level* is based on *L1Id*, *L2Id*, *L3Id* and *L4Id* looked up in the *UserReqLevelName* table.

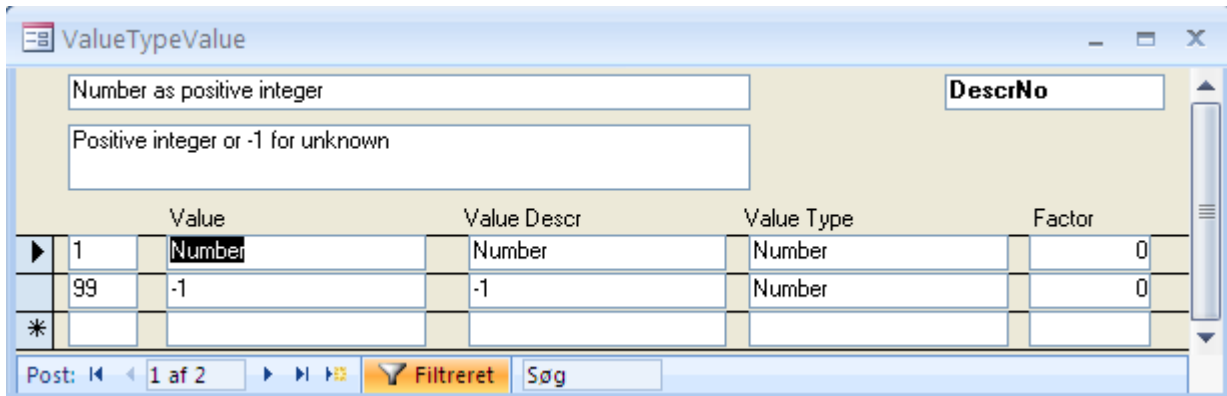
Click on command buttons *View calculation instructions* will open the form shown in Figure 39 'The requirements form for specification of calculations of user requirements' which gives the description of the function to calculate the value. The form is given the related *UserReqId*.

II.5 Miscellaneous

This section lists the various miscellaneous forms in the BR-ReMS.

II.5.1 Value type descriptions

All forms showing value types of characteristics can possibly have command buttons leading to the form 'Value type form with description of possible values for a type' shown in Figure 59.



The screenshot shows a window titled 'ValueTypeValue'. It contains two text input fields: 'Number as positive integer' and 'Positive integer or -1 for unknown'. To the right is a 'DescrNo' field. Below these is a table with the following columns: Value, Value Descr, Value Type, and Factor.

	Value	Value Descr	Value Type	Factor
▶	1	Number	Number	0
	99	-1	Number	0
*				

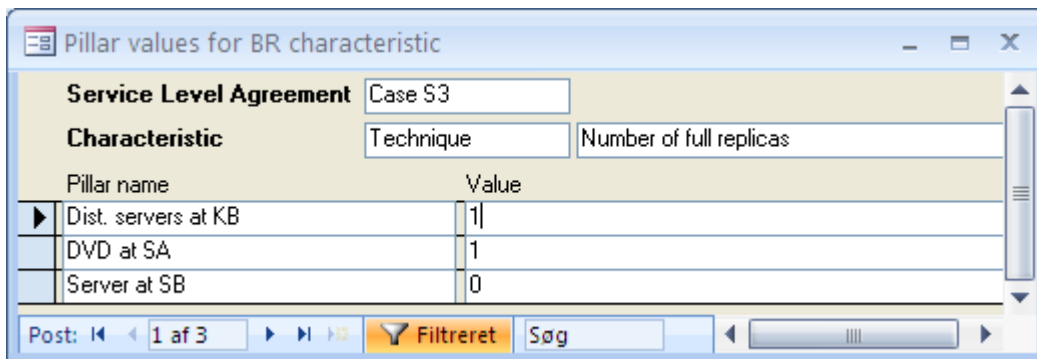
At the bottom, there is a navigation bar with 'Post: 1 af 2', a 'Filtreret' button, and a search field labeled 'Søg'.

Figure 59 Value type form with description of possible values for a type

A value type shown by the form represents an entry in the *ValueType* table.

II.5.2 Overview of all pillar values for one characteristic

Help form to show values for a pillar characteristic for all pillars in a SLA is given in form 'Overview of all pillar values for one pillar characteristic and pillars in a SLA' shown in Figure 60.



The screenshot shows a window titled 'Pillar values for BR characteristic'. It contains two text input fields: 'Service Level Agreement' with the value 'Case S3' and 'Characteristic' with the value 'Technique' and 'Number of full replicas'. Below these is a table with the following columns: Pillar name and Value.

Pillar name	Value
▶ Dist. servers at KB	1
DVD at SA	1
Server at SB	0

At the bottom, there is a navigation bar with 'Post: 1 af 3', a 'Filtreret' button, and a search field labeled 'Søg'.

Figure 60 Overview of all pillar values for one pillar characteristic and pillars in a SLA

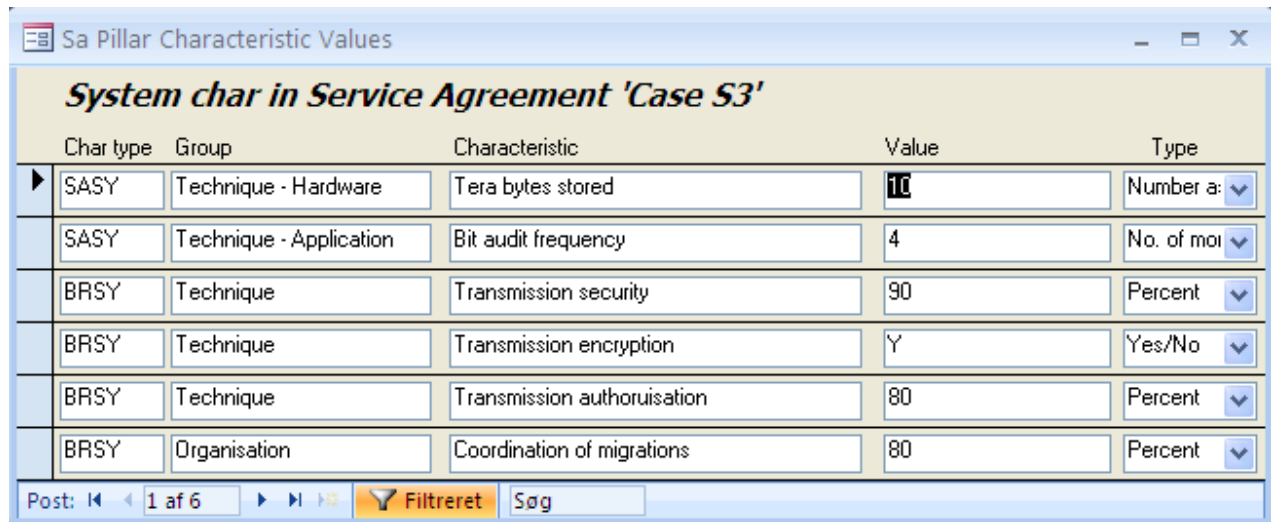
The *Service Level Agreement* field shows information about the given *Slaid* based on information from the *Slat* table.

The *Characteristic* field shows information about the given *CharId* based on information from the *Char* table.

The list *Pillar Name* shows the pillars chosen for the SLA, and the related *Value* is the *Value* from the *FctResPillarCharValuetable* table, the *SlapillarCharVal* table or the *BrPillarCharVal* table, depending on what type of characteristic that was given.

II.5.3 Overview of all systems characteristics

Help form to show values for all system characteristics is given in form 'Overview of all systems characteristics' shown in Figure 61.



The screenshot shows a window titled "Sa Pillar Characteristic Values" with a header "System char in Service Agreement 'Case S3'". Below the header is a table with the following columns: Char type, Group, Characteristic, Value, and Type. The table contains six rows of data. At the bottom of the window, there is a navigation bar with "Post: 1 af 6", a "Filtreret" button, and a search box labeled "Søg".

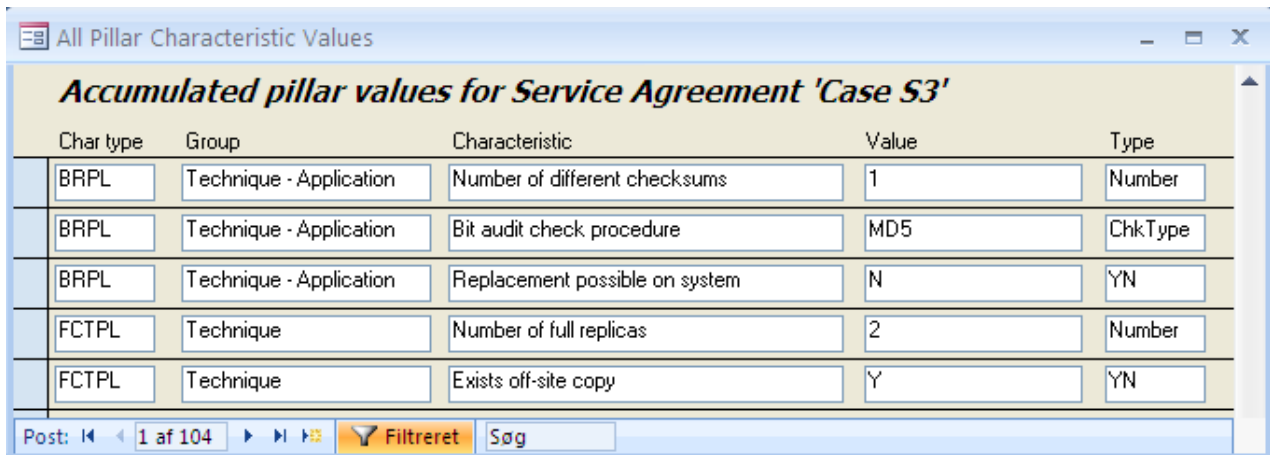
Char type	Group	Characteristic	Value	Type
SASY	Technique - Hardware	Tera bytes stored	10	Number a: ▾
SASY	Technique - Application	Bit audit frequency	4	No. of mor ▾
BRSY	Technique	Transmission security	90	Percent ▾
BRSY	Technique	Transmission encryption	Y	Yes/No ▾
BRSY	Technique	Transmission authorisation	80	Percent ▾
BRSY	Organisation	Coordination of migrations	80	Percent ▾

Figure 61 Overview of all systems characteristics

The list values are based on the *FctPillarCharAccValue* table, the *SlaSysCharVal* table and the *BrSysCharVal* table. Each line represents a characteristic with characteristic information based on the Char table, where *Group* is based on *L1Id*, *L2Id* and *L3Id* looked up in the *CharLevelName* table, and the *Type* is looked up in the *ValueType* table.

II.5.4 Overview of all pillar characteristics

Help form to show values for all pillar characteristics is given in form 'Overview of all pillar characteristics' shown in Figure 62.



Char type	Group	Characteristic	Value	Type
BRPL	Technique - Application	Number of different checksums	1	Number
BRPL	Technique - Application	Bit audit check procedure	MD5	ChkType
BRPL	Technique - Application	Replacement possible on system	N	YN
FCTPL	Technique	Number of full replicas	2	Number
FCTPL	Technique	Exists off-site copy	Y	YN

Figure 62 Overview of all pillar characteristics

The list values are based on the *FctResPillarCharValuetable* table, the *SlaPillarCharVal* table and the *BrPillarCharVal* table. Each line represents a characteristic with characteristic information based on the *Char* table, where *Group* is based on *L1Id*, *L2Id* and *L3Id* looked up in the *CharLevelName* table, and the *Type* is looked up in the *ValueType* table.

II.5.5 Queries

There are a number of queries in the database, where some of them produce related data in the tables. For instance, if a new entry is created in the *Char* table with type "BRPL", then there must also be created related entries in the *BrPillarCharVal* etc.

II.5.6 Reports

There are a number of reports in the database based on data in the tables. For instance, a report to give an overview of the different types characteristics defined in the BR-ReMS.

Appendix III. The BR-ReMS Data Model

This appendix contains the BR-ReMS Data Model. The below Figure 63 shows the full BR-ReMS data model. The data model is illustrated in an Entity-Relationship diagram using crowfoot notation.

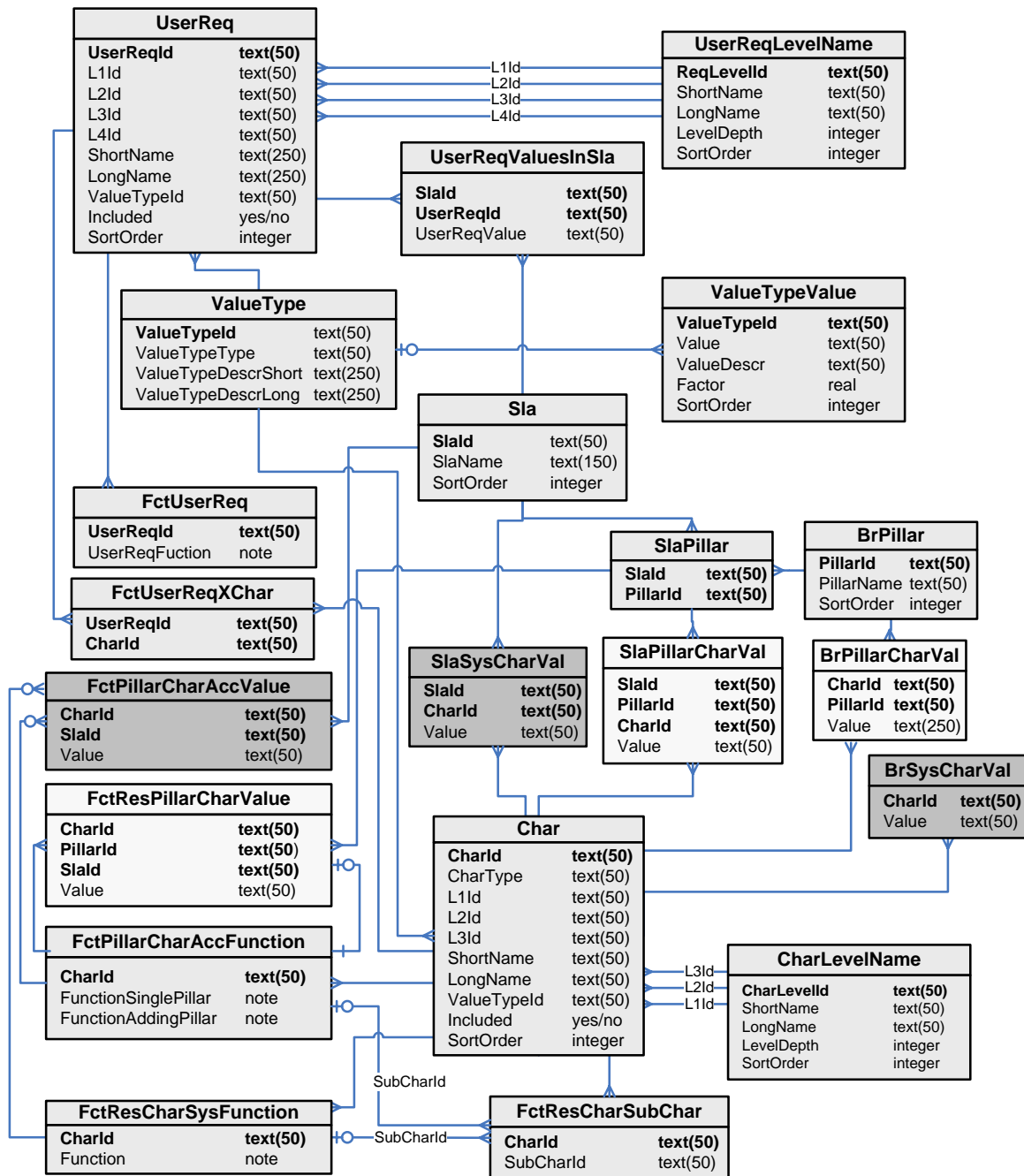


Figure 63 The BR-ReMS data model

Note that:

- Attributes in bold are primary keys for an entity
- Attribute names are the same in relations, except when other name is written on relation.
- Light grey entities concern pillar level
- Dark grey entities concern system level

Appendix IV. An example of a WARC files

This appendix gives an example of the *WARC repr* file and for the *WARC p1* file from Figure 27 based on the METS given in Figure 26. In the given examples, there will be used italic font for information that is describing what the contents would be in a practical example. It should also be noted that these examples only presents a very limited use of the WARC format.

The *WARC repr* file has the contents:

```
WARC/1.0
WARC-Type: warcinfo
WARC-Date: 2011-08-08T12:30:00Z
WARC-Record-ID: <urn:uuid:8d39797c-7e4c-4110-baa5-e44a8a725276>
Content-Length: 115
Description of the contents of the WARC-file in form of fields. This is not
restricted to any specific information.

WARC/1.0
WARC-Type: resource
WARC-Date: 2011-03-28T12:30:00Z
WARC-Record-ID: <urn:uuid:ezTIFFSmdUUID>
Content-Length: 1553
<mets OBJID="urn:uuid:ezTIFFSmdUUID">
  <amdSec>
    <techMD CREATED="2011-08-014T12:10:00">
      <mdWrap MDTYPE="PREMIS:OBJECT">
        <xmlData>
          <object xsi:type="representation">
            <linkingIntellectualEntityIdentifier>
              <linkingIntellectualEntityIdentifierType>
                URN
              </linkingIntellectualEntityIdentifierType>
              <linkingIntellectualEntityIdentifierValue>
                urn:uuid:LOGICEzUUID
              </linkingIntellectualEntityIdentifierValue>
            </linkingIntellectualEntityIdentifier>
          </object>
        </xmlData>
      </mdWrap>
    </techMD>
    <digiprovMD CREATED="2011-08-01T12:10:00">
      <mdWrap MDTYPE="PREMIS">
        <xmlData>
          <preservationLevel>
            <preservationLevelValue>
              HighBitSafety
            </preservationLevelValue>
          </preservationLevel>
        </xmlData>
      </mdWrap>
    </digiprovMD>
  </amdSec>
  <structMap>
    <div>
      <mptr ID="urn:uuid:ez01mdUUID" LOCTYPE="URN"/>
    </div>
    <div>
      <mptr ID="urn:uuid:ez02mdUUID" LOCTYPE="URN"/>
    </div>
  </structMap>
</mets>
```


The WARC-record containing the actual TIFF p1 file will have the identifier:

[urn:uuid:ez01srcUUID](#)

It is worth noting that the bit repository will get a BR SIP as WARC package with an identifier for the WARC package. That means that computation is needed in order to resolve the identifiers for the TIFF, TEI-P4 and METS files. This computation can be placed either in the IR or BR. The important point is that either way the computation needed is only relying on the WARC format, since an index of all identifiers can be produced on bases of the bit preserved data.