On the Need for Benchmark Corpora in Digital Preservation

Robert Neumayer[1], Hannes Kulovits[1], Andreas Rauber[1],
Manfred Thaller[2], Eleonora Nicchiarelli[3], Michael Day[4], Hans Hofman[5], Seamus Ross[6]

1) Vienna University of Technology
{neumayer,kulovits,rauber}@ifs.tuwien.ac.at
2) University of Cologne
manfred.thaller@uni-koeln.de
3) Austrian National Library
eleonora.nicchiarelli@onb.ac.at
4) University of Bath
m.day@ukoln.ac.uk
5) National Archives of the Netherlands
hans.hofman@nationaalarchief.nl
6) University of Glasgow
s.ross@hatii.arts.gla.ac.uk

Abstract

The main purpose of this paper is to clarify a range of issues arising in the context of building an open testbed digital object corpus for digital preservation research. A corpus should be substantial and consistently designed as well as include digital objects belonging to different genres, having different technical characteristics and appearing in different contexts. For the means of ensuring maximum benefit from such a corpus, the main challenges in this context are the practicability of an open corpus as well as the combination of requirements. The unique challenges like the huge complexity and novel goal in the area of corpus building for digital preservation will also be addressed.

Categories and Subject Descriptors

H.3 [Information Storage and Retrieval]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries; H.2.3 [Database Managment]: Languages—*Quer*y *Language*s

General Terms

Measurement, Performance, Experimentation

Keywords

Corpora Building, Benchmark Corpora, Digital Preservation

1 Introduction

Comparability of results is a key issue in many disciplines and has proven to be extremely helpful for both collaboration and competition amongst research groups. Benchmarking corpora are clearly essential in order to devise well-defined tasks on which researchers can test their algorithms. In this context, we can define a corpus as an annotated collection of files, documents, or digital objects, where the annotations represent the criteria the algorithms will be evaluated against. The e-mail filtering domain, for instance, provides corpora containing a number of e-mail messages labelled as either ‘legitimate’ or ‘non-legitimate’ (the latter also known as spam), on which researchers can test their own filtering implementation against the manual classification (ground truth), yielding a quantitative ranking of filtering algorithms. This example shows that corpora can be of great value if they are well-defined and well-known within the community.

The need for a thorough evaluation of options for corpora creation in the domain of digital preservation is not only driven by its complexity (compared to the e-mail filtering example). The emerging importance of digital preservation issues along with its impact on the research landscape makes clear the necessity for a suitable means of comparing results amongst researchers across institutional backgrounds and domains. Unlike 'traditional benchmarking corpora' areas like machine learning or information retrieval, the scope of digital preservation reaches far beyond the area of computer science, bringing together requirements from many stakeholders with different backgrounds. As many participants contribute to digital preservation research, many scenarios and requirements are vital building blocks of requirement gathering in this context. The Delos Network of Excellence work package 6, task 6.9: 'Development of an Open Testbed Document Corpus' focuses on corpus development for digital preservation [8].

Having the purpose of giving recommendations and guidelines for creating a benchmark corpus for digital preservation, this paper starts by giving a short overview of the topic. We continue by giving examples of existing benchmark corpora and key factors and merits for the concerned communities. Further, objects will be categorised to allow easier handling of stratification issues, i.e. ensuring that all essential groups of digital objects are represented according to their relevance. Further, we talk about involved legal aspects. Finally, we give a short overview of other projects that could possibly profit from digital preservation benchmarking corpora of digital objects, and list open issues.

2 Creation of Benchmark Corpora

In every case, building a benchmark corpus is a difficult task. Generally, at least the following aspects of system evaluation are relevant when creating benchmark corpora, and have thus to be taken into account:

System quality evaluation: the measurement of the correctness of applied methods or algorithms with respect to the quality of the outcome. Quality could be assessed by comparing the results given by a tool to those given by another, widely accepted standard tool or method, which of course, would have to be identified first.
System performance evaluation: it should cover both performance per instance/object, and scalability issues (i.e. how the performance changes when the methods are applied to large collections).

This means that the collections themselves that are part of the benchmark have to satisfy certain quality constraints. If the application of machine-learning algorithms is foreseen, digital objects in the corpus have to be tagged, or contain sufficient metadata. The available range of digital objects should be as wide as possible, in order for the collection to be close to one or a set of real-world scenarios. If it is intended to perform scalability experiments, a suitable size has to be guaranteed, where the appropriate number of objects should be estimated according to the requirements of the use cases for the application to be tested. In the digital preservation context this is of particular importance, since most of the applications will very likely have to be applied to huge numbers of digital objects and a broad range of different institutional backgrounds. In general, corpora can be categorised by their purpose: a digital object corpus can be 'content-complete' or 'feature-complete'.

We introduce the term object type, denoting a certain type of files like, e.g. text documents. This layer of abstraction is situated above the layer of actual file types. The text document type, for instance, would comprise file types like .doc or .sxw, i.e. Microsoft Word and OpenOffice writer formats.

Content-complete corpus:

A 'content-complete' corpus covers the widest variety of possible types of content available in a given scenario. The purpose of such a corpus is the testing of organisational procedures: for instance, all the digital object types used in a given organisation or setting are covered, in order to describe their specific problem definitions.

Feature-complete corpus:

A 'feature-complete' corpus is defined by the coverage of the widest variety of possible features of a given object type. Such a corpus therefore is, by definition, object-type specific, and can be used to test the completeness of implementations for a given object type.

Performance-defining corpus:

A set of objects which is sufficiently large so that two or more programs processing it can be compared in a meaningful way with respect to their performance. A combination of both 'content-' and 'feature-complete' corpora would make a good choice for a performance-defining corpus. In a simpler scenario, for instance for performance measurements for a specific object type, a 'feature-complete' corpus could be simply created by injecting the same object(s) multiple times (if the only purpose is testing the available algorithms for scalability).

In the area of programming languages, we usually do not speak of corpora of problems, but of benchmark tasks. The reason for this is, that in defining a benchmark the assumption of the previous paragraph, that 'passing a benchmark' guarantees a well defined quality, is central for software engineering; in linguistics, from which the concept of a corpus is originally derived, that assumption is much weaker, as the understanding that even very large corpora are only a subset of all possible linguistic expressions is more central. Within these disciplines the purpose of a corpus is primarily the creation of a well defined frame of reference for the scientific discourse.

When we talk about corpora of digital objects – particularly of objects used for preservation purposes – this difference between a 'benchmark', as defined within software technology, and a 'corpus', as derived from linguistics, becomes blurred, which is why we use the term 'benchmark corpus' in this paper. Some of the reasons for merging the two concepts are:

1. More and more digital objects are essentially active programs, which direct the operations of a hardware system, albeit they usually do so through an interpretative layer, as opposed to static containers of data, which are interpreted by other software systems. A postscript file is the sourcecode of a program written in the Postscript language. The scene graph of a VR application, most obviously in the case of VRML, is a program within a programming language.
2. At the same time, however, the discussion of preservation has reached sufficiently far beyond engineering terms, that complete preservation is as impossible as a complete translation of a natural language expression. The lossless transfer of a linguistic expression from an English syntax graph into a, say, Italian one is possible. A correct idiomatic, or even a full semantic translation from an English into an Italian utterance is also possible. To succeed with both at the same time is not. Similarly the preservation of all the Shannonian information contained within one digital object into another object is possible. The preservation of the associations evoked by a digital system by its behaviour[1] at a specific time may be preservable. Both types of preservation at the same time might be contradictory within themselves.

Feature complete benchmark corpora therefore occur in three flavours:

[1] It may be thought of a system which directs the focus of a user to a specific item, by emphasising it by 'a highly unusual set of attributes'. This is a behavioural feature, which depends on the socio-technical context in which the system is executed, which can be preserved – but by definition only, by changing its technical realisation more or less permanently.

- Engineering feature complete corpora provide examples of a complete list of technical features which can occur within a certain class of digital objects (e.g. sound files).
- Discourse feature complete corpora define a set of properties, which are recognised and valid for a specific content oriented discussion.
- Policy feature complete corpora give examples for all features digital objects may have, which are supported by a digital (preservation) policy.

All types of feature complete corpora define norms derived from some set of abstract rules. In that respect they are different from content complete corpora, which are derived from the empirical observations of the occurrence of object types within an environment.

The possible benefits of benchmarking corpora for the information retrieval community are pointed out in [3]: in this article guidelines to avoid known problems faced by past collections are given, the most relevant issues mentioned being 1) needs concerning sets of documents, 2) needs concerning individual documents, and 3) needs concerning relevance judgements. Altogether this should make sure that relevant documents are used, collections always have a well-defined purpose, and the relevance judgements or evaluation criteria are clear. All three points are also applicable to corpora for digital preservation. We will give an overview of existing corpora – amongst others in the information retrieval domain – in the next section.

3 Existing Benchmarking Corpora in Information Retrieval and Machine Learning

Across various disciplines many benchmarking corpora have been created. Some of them 'happened' rather than they were actually designed or planned, either for lack of (public) alternatives or their sheer popularity and availability. In this section, we will give a short overview of existing benchmarking corpora, their merits and problems, as well as related open issues.

3.1 TREC - Text Retrieval Conference

In information retrieval the TREC conference and its benchmark corpora are widely used [5]. Every year, information retrieval tasks in different domains, are offered, also known as 'Tracks'. TREC therefore covers a wide range of general information retrieval tasks, including video or text classification, performance tasks, or question answering.

Another important issue is sustainability, which is sometimes achieved by distributing the corpus for money only. The TREC web research collections, for example, are hosted by the University of Glasgow and are distributed for fees from 250 up to 600 UK Pounds and have a size from two to 426 gigabytes.

3.2 The MIREX Audio Corpora

The domain of music information retrieval faces huge problems when it comes to publicly available corpora. The copyright situation for audio material is a very strict one. Once a year the International Conference of Music Information Retrieval (ISMIR) hosts an evaluation contest to provide researchers a chance to compare their algorithms [4]. The goal of this contest is to compare state-of-the-art algorithms and systems relevant for Music Information Retrieval.

Due to copyright restrictions on the audio files which were used to run the experiments, submissions had to be sent in following strict code guidelines so that the actual test runs would run smoothly on the site hosting the benchmark corpus. No details of the corpus were released to the participants. Results were published on anonymised file IDs only, rather than listing any music titles. In this year, two tasks, Audio Music Similarity and Retrieval and Symbolic Melodic Similarity required post-run human evaluations. A lot of effort has been put into the preparation and evaluation of this contest, yet it made use of one object type only.

3.3 UCI - Machine Learning Repository

Another popular repository for benchmarking corpora is the UCI Machine Learning Repository [7]. The corpora offered cover a wide range of different domains (housing, finance, medical) and some of them have restricted access only, e.g. medical data sets. The UCI repository contains both artificial, e.g. a zoological database describing seven types of animals, as well as real-world data like medical records. Data sets differ in terms of missing values, the type of learning problem they require (e.g. classification or numeric prediction), or the types of values occurring (e.g. binary or nominal values). Moreover, the range of available classes is quite large. Prominent data sets within that repository are the 'Boston Housing Database', concerning housing prices in the suburbs of Boston or the 'Pima Indians Diabetes Database', covering information about Indians tested positive or negative for diabetes.

The data sets in the UCI repository are mostly used for the most common machine learning tasks classification and clustering, the categorisation of similar instances into homogeneous groups. The number of instances varies in between datasets, i.e. the collections also differ in size.

4 Challenges for Digital Preservation Corpora

Taking all this into consideration we identify five great challenges for corpus generation in the digital preservation context. These challenges are shaped in appreciation of the effort made in 40 years of corpus building and the experience as well as benefits it has brought to its communities.

- Precise Task Definition
- Size
- Stratification
- Data Representation
- Ground Truth and Evaluation Criteria

4.1 Precise Task Definitions

More than anything else, a thoroughly defined, precise task definition is vital for the success/usefulness of a corpus for digital preservation. As can be seen in the corpora examples given above, all corpora are highly domain- and task-specific (e.g. e-mail categorisation into 'legitimate' and 'non-legitimate' messages). Many parameters or goals strongly depend on the application scenario, i.e. an archival setting will probably put more emphasis on authenticity than computer science research teams.

4.2 Size

The size of a corpus either means the number of objects involved or the actual file or corpora sizes in terms of hard disk space needed. 'Sufficient' size always depends on the task and the chosen stratification strategy. 'Feature complete' corpora have clearly different requirements than, for example, 'content complete' corpora, as very few files may cover all features of an object type (at least for simple types). Tasks concentrating on scalability issues will also require a substantially larger number of test instances. Special object types that are not very common or available online (e.g. copy protected, encrypted or restricted object types) should be considered especially.

4.3 Stratification

Stratification denotes the coverage of all necessary types of digital objects, domains, and varieties thereof which are required by the community in a specific context, i.e. the distribution of elements in a corpus according to different criteria.

For the digital preservation context possible stratification categories could be:

- File type: For instance, if image types were underrepresented in the result of random sampling, more image files could manually be injected, guaranteeing the correct level of stratification.
- Categories/Classes: Files in benchmark collections should be as diverse as possible in terms of classes or categories, i.e. a benchmark corpus needs to cover an appropriately wide variety of classes.
- Time: Object types and specifications change a lot over time, this is also closely related to versions of file types. Especially for long term preservation it is a crucial task to include elements from different periods of time.
- Real-World scenario: To be as close to a real world scenario and achieve reasonable size, using a web archive seems to be the best option, the possible inputs for that would be a certain time span or domains, the rest of variety should emerge from the random sampling.

Stratification can be performed according to different stratification needs, aiming for example either at an equal distribution of data items across all categories present in the corpus, or aiming at a distribution reflecting real-world settings.

4.4 Data Representation

All complex problems need to be transformed to a machine-readable and processable form. The general idea is to represent complex digital objects like documents or facts about the real world in a computer-readable form, making it essential to compare results. Transforming complex facts into vectors or matrices is a challenge itself. The coding of nominal attributes (unordered categories) and ordinal attributes (order is important), for example, must be taken into account. Version numbers or identifiers for the PDF standard for example can take values of '1.5, 1.4, 1.3', and hence are ordinal data since their order is important. Other possible types of attributes are boolean (either true or false, i.e. existing of missing), continuous values (e.g. natural or floating point numbers), or even textual data.

Various types of transformations or look-up tables are feasible to present data in a machine-readable and meaningful way. The complexity of the problem rises as more compound or sophisticated objects shall be represented.

4.5 Ground Truth

Ground truth denotes the criteria to be evaluated against, often determined by human evaluators. Their annotations are then used to judge the correctness of systems. For every corpus the definition of a ground truth is essential. It has to be made clear what the precise use of such a labelling will be. For the text categorisation task, the ground truth is defined as each digital object's class value, which is exactly the criteria for evaluation. Ground truth can be defined by human annotators, raising, for example, problems of inter indexer consistency (i.e. do annotators provide consistent annotations or not). Especially the audio similarity contests described before required a lot of effort for annotations.

5 Categorisation of Features and Most Relevant Digital Object Types

A comprehensive overview of file format and object types is given in the 'File formats typology and registries for digital preservation' deliverable of the DELOS project [2].

Files or digital objects can be organised in categories or content classes, i.e. groups of files that have common properties. For certain scenarios, a balanced corpus containing files from a number of categories in sufficient size is relevant. One content class therefore can comprise several object types, e.g. the content category text document will comprise .doc, .rtf, .pdf files etc. The next step towards an open digital object benchmark corpus is the identification of existing file types and their specific features with respect to the most common object types. Several types of objects are presented in [1].

A categorisation like this can be used for stratification purposes. An image corpus, for instance, should comprise files from both the 'vector-based' and the 'pixel-based' categories. A typology makes it easier to check whether certain file types are under or overrepresented. Such a categorisation can also help to identify objects that have similar properties or essential characteristics. Subsequently, processing functionality can be implemented hierarchically as well. For instance, if the object in question is a *.rar* file, the tree is traversed down to the correct entry in the *archive* category. The next logical step would be to decompress the archive and start traversing the tree for all extracted files. Embedded images would be treated analogously. Those categories can also help to create a corpus that contains objects well distributed across these types (i.e. create a uniformly distributed corpus according to second level categories).

Subsequently it makes sense to identify relevant features, also referred to as 'significant properties' or 'significant characteristics' and assign them to those categories. The first level of properties to be considered are the most general types of information about any file (note that these include not only technical file format characteristics, but also include conceptual properties such as relationships between files, for e.g. authenticity requirements):

- Application used for creation (and viewers)
- Region of creation (locales, domains, languages, character sets)
- Version of file format
- MIME type or automatic identification of MIME-types may still be of interest for categorisation purposes, it has to be differentiated between a file's MIME-type and its actual content.
- Year or creation date
- Genre
- Container (whether the digital object includes or references other documents)
- Technical characteristics specific to a given object type

Of course, this list can and should be extended as other object types are identified to be relevant as well as updated by inputs from content partners specifying their requirements. The Data Dictionary for Preservation Metadata by the PREMIS initiative specifies metadata for digital preservation, which may constitute an essential set of attributes to provide within a digital preservation benchmark corpus [9]. Yet, note that PREMIS specifically does not elaborate technical and hardware specific metadata. Some technical properties for some object types can and will automatically be extracted by characterisation tools and subsequently automatically validated. Of course, tools will be needed for automatic extraction of file characteristics; there do exist a number of characterisation tools and some may be further developed.

Table 1: Exemplary, simplified metadata to be covered for different filetypes.

Attribute	PDF	PNG	MP3	AVI
Version	+	+	-	+
Encoding	+	-	-	-
MIME	+	+	+	+
Created by Application	+	-	+	+
User-Defined Metadata	+	+	+	-
Year or Creation Date	+	+	+	+
Composite	+	-	-	+
Images	+	-	-	-
Audio	+	-	-	+
Video	+	-	-	+
3D Elements	+	-	-	-
Subtitles	-	-	-	+
Genre	+	+	+	+
Compressed	-	+	+	+
Resolution	-	+	-	+
Sample Rate	-	-	+	+
Codec	-	-	+	+
…	…	…	…	…
∞	∞	∞	∞	∞

As an example for the different levels, which characteristics can take, we give the following list of common characteristics of images:

Low level characteristics

are needed to mechanically process the image data. An image may be encoded with Little / Big Endian encoding. The compression type used is of great interest. It is also important to know about the interlacing mode of an object.

Format characteristics

are used to describe the image objects like their size (width, height) or colour depth.

Characteristics needed to preserve the visual interpretation

are, for instance, the photo-grammetric interpretation or an image's colour profile.

Characteristics needed to preserve image behaviour

Information about, for example, embedded links between subimages might be of great help.

The above characteristics can be automatically validated via characterisation tools, i.e., if they are wrong the processing of the object is ideally not endangered. The characteristics called 'arbitrary' here can be inherently wrong, without hindering the processing of the object (e.g. wrong annotations). However, these can not be automatically validated.

Characteristics needed to preserve the semantic interpretation:

- Arbitrary engineering metadata (equipment type)

- Arbitrary semantic metadata (creator, content, copyright)

Table 1 shows a simplified attempt to cover file characteristics for four file types. However, it makes clear that an exhaustive enumeration of file characteristics is very likely to reach an extremely high number, leading to a virtually unlimited number of potential corpora or stratification criteria. Therefore, benchmarking corpora for digital preservation need to be highly task-specific. The table implies, for instance, that a scenario that covers only the basic media types could quickly run out of hand. Comprehensive characteristics are covered in various standardisation documents, e.g. the PDF standard. The PREMIS data dictionary for preservation metadata extensively lists metadata elements for a wide range of digital objects [9].

6 Populating a Benchmark Corpus and Collaboration Aspects

Having pointed out the main problems and most difficult issues, a possible strategy for corpora creation and population is outlined below:

1. Requirements Analysis: is used to gather requirements for possible corpora building. This includes thorough investigation of key problems and needs of the particular institution or domain. This step should include either all people from a specific institution if they want to test their own implementations or explore solutions for their own data or all or as many people as possible from their domain, e.g. all scientific partners of the project in the case of a corpus for academic publications. The output of this step should be a precise definition of the object types to be included and a specific list of characteristics to be taken into account. A starting point for this might be the case studies on preservation planning that have been undertaken as part of the DELOS Testbed and the PLANETS project [10, 11].
2. Define Precise Tasks: All possible purposes of the corpus have to be prescribed. There will be experiments possible that are not stipulated at this point in time. However, repurposing does not guarantee optimal outcomes. Each corpus has to pose a certain value for the community and is therefore designed for a specific, well-defined task.
3. Acquire Data: Once the decision for a particular object type or set thereof is made, consensus has to be reached about the size of the corpus and stratification issues as well as the actual files that will be used for the corpus and subsequent experiments.
4. Define Ground Truth: This step includes the definition of experiments to be performed on the corpus. All annotations for files in the corpus have to be decided here as well as how these annotations will be made.

   Defining a ground truth for digital preservation corpora is inherently difficult and often even impossible because the question which preservation solution is the ideal one always depends on the specific context and the requirements of each institution. Instead of defining one ground truth, it would be possible to provide a set of different ground truths for specific purposes like retrieval or classification. Moreover, different institutions might have fundamentally different requirements, which could be reflected in such a way. The DELOS testbed activities, continued in the PLANETS project, have included performing a series of case studies aiming at eliciting the requirements for specific ground truth definitions.
5. Define Evaluation Criteria and Procedure: Finally, measures have to be defined on how to evaluate the tasks. How is the evaluation to be done? Predefined splits between training and validation set have to be agreed on. What are the requirements for participants to be successful?

7 Legal Issues

A range of legal issues might be faced during the building of a benchmark corpus similar to the one described in this paper. This section only points out very few important questions, a more extensive discussion is given in [6].

The clarification of legal issues for an internationally contributed corpus made available on the internet is a long and complex process, that should be started as soon as possible.

The general discussion about legal issues should be strongly influenced by who is going to host a possible corpus – and where. The first step should be understanding whether the legislation of the country where the corpus is hosted is applicable in case of copyright infringement for vast differences exist across countries. Subsequently, whether the corpus has uniquely scientific purposes, and copyright transfer issues should be investigated. All matters covered by national law, as privacy, security, and data protection, should be separately investigated.

8 Outlook on Collaboration Possibilities

A number of projects are active in the area of digital preservation, and therefore are potential partners in building benchmarking corpora. Some projects also have at least some effort designated to benchmarking. To guarantee maximum value for all member projects, particularly content providers, collaboration plays a key role in corpus building. Content partners are required to propose object types they have a particular interest in, that could be either proprietary or not widespread available yet (i.e. rather new formats that will be essential in the future). Participation could happen via registration schemas for corpora submission or public discussion about needed types of corpora. Collaborative requirements definition could vastly improve the overall value corpora pose to the community as a whole.

Characterisation tool developers could provide vital inputs for evaluation and feature extraction possibilities, covered in Section 5 (which object types they concentrate on, which features they extract).

Possible inputs could be:

- Files of different types
- Metadata
- Descriptions according to Section 5
- Type of corpus information (content complete, feature complete for PDF)

The PLANETS project[2] is working on implementing a digital preservation testbed, having a strong need for data collections testing their implementation for characterisation as well as migration tools.

Acknowledgements

Part of this work was supported by the European Union in the 6th Framework Program, IST, through the DELOS NoE on Digital Libraries, contract 507618.

References

[1] Lars R. Clausen. Handling file formats. Technical report, The State and University Library, Århus, Denmark, The Royal Library, Copenhagen, Denmark, 2004.

[2] Maria Guercio and Cinzia Cappiello. File formats typology and registries for digital preservation. Technical report, DELOS - Network of Excellence on Digital Libraries, 2004.

[3] Karen Spärck Jones and Cornelis Joost van Rijsbergen. Information retrieval test collections. *Journal of Documentation*, 32:59–75, 1976.

[4] Annual Music Information Retrieval Evaluation eXchange (MIREX). Website, 2005. `http://www.music-ir.org/mirexwiki/index.php/Main_Page`.

[5] National Institute of Standards and Technology. Text retrieval conference (TREC). `http://trec.nist.gov/`, accessed January 8, 2007.

[2] `http://www.planets-project.eu`

[6] Robert Neumayer, Christoph Becker, Thomas Lidy, Andreas Rauber, Eleonora Nicchiarelli, Manfred Thaller, Michael Day, Hans Hofman, and Seamus Ross. Development of an open testbed digital object corpus. Technical report, March 2007.

[7] David J. Newman, Seth Hettich, C.L. Blake, and Christopher J. Merz. UCI repository of machine learning databases, 1998.

[8] Delos NoE. Detailed joint programme of activities - third period, June 2006.

[9] Preservation Metadata Implementation Strategies Working Group (PREMIS). Data dictionary for preservation metadata. Final report of the premis working group, Online Computer Library Centre (OCLC) and RLG, May 2005.

[10] Stephan Strodl, Christoph Becker, Robert Neumayer, Andreas Rauber, Eleonora Nicchiarelli Bettelli, Max Kaiser, Hans Hofman, Heike Neuroth, Stefan Strathmann, Franca Debole, and Giuseppe Amato. Evaluating preservation strategies for electronic theses and dissertations. In *Working Notes of the DELOS Conference*, pages 297–305, Pisa, Italy, February 13-14 2007.

[11] Stephan Strodl, Andreas Rauber, Carl Rauch, Hans Hofman, Franca Debole, and Giuseppe Amato. The DELOS testbed for choosing a digital preservation strategy. In *Proceedings of the 9th International Conference on Asian Digital Libraries (ICADL'06)*, pages 323–332, Kyoto, Japan, November 27-30 2006. Springer.