



KB and migration
Working document

Version: 0.2
Author: Caroline van Wijk
Date: 7 August 2006

National Library of the Netherlands (Koninklijke Bibliotheek)
Digital Preservation Department

KB and migration

History of KB and migration

Document version	Date of modification	Author	Summary of modifications
0.1	4 July 2006	Caroline van Wijk	
0.2	7 August 2006	Caroline van Wijk	Remarks H. van Wijngaarden and J.Rog incorporated. <ul style="list-style-type: none">• Added: characterisation of digital object on functional and technical base.• Extended: normalisation is an option for all uncommon file formats, the files from the DARE project are a concrete example.• Paragraph 2.3 Characterising digital objects has been altered.

Related documents

Project Initiatie Document – Migratieonderzoek, C. van Wijk (2006)

Starting Point Migration Research, C. van Wijk (2006)

Test plan Migration Research, C. van Wijk (2006)

KB and migration

Table of Content

1	Introduction	4
1.1	Purpose of this document	4
1.2	Sections in this document	4
	<i>Content e-Depot</i>	<i>4</i>
	<i>Possibilities and consequences of using migration at the KB.....</i>	<i>4</i>
	<i>Appendix content e-Depot.....</i>	<i>5</i>
2	Content e-Depot.....	6
2.1	Introduction	6
2.2	Archived objects in the e-Depot	6
2.3	Characterisation of digital objects	6
3	Possibilities and consequences of the use of migration at the KB.....	8
3.1	Introduction	8
3.2	Results File format research project	8
3.3	Test Migration research project	8
3.4	Application of migration at the KB	9
I	Appendix content e-Depot.....	10
I.1	Example of functional characterisation of an e-Depot article.....	11
	<i>Content</i>	<i>12</i>
	<i>Structure.....</i>	<i>14</i>
	<i>Appearance</i>	<i>16</i>
	<i>Context</i>	<i>20</i>
	<i>Behaviour.....</i>	<i>21</i>
1.2	Example technical characterisation of e-Depot article	22
1.3	Overview (preliminary) functional characterisation of e-Depot content.....	25

1 Introduction

1.1 Purpose of this document

The current (summer 2006) possibilities for the use of migration as a digital preservation strategy at the National Library of the Netherlands are discussed in this working document.

This working document contains an overview of the possible application of variations of migrations (discussed in *Starting Point Migration Research*). The overview of application is based on the (future) content of the e-Depot and the results of the File format research project. The File format research project is still in progress. Future results of this project will be incorporated in this document.

This document is used as a starting point for the test plan for migration tools.

Additional remarks and suggestions are welcome and can be sent to the author (caroline.vanwijk@kb.nl).

1.2 Sections in this document

Content e-Depot

This chapter contains overviews of the types of objects that have been and/or will be archived in the e-Depot.

Moreover, the chapter contains examples of a functional and technical characterisation of the most common type of object in the e-Depot.

This characterisation can be used as a framework for migrating objects archived in the e-Depot. It is necessary to define the essence of the original object to be able to conclude whether a conversion result is correct. A functional characterisation is possible by using the five properties of a digital object: content, structure, appearance, behaviour and context. A technical characterisation is dependent on the file format of the specific object.

Possibilities and consequences of using migration at the KB

In this part of the document the possibilities and consequences of the use of migration, based on the content of the e-Depot and the known variations of migration, are described.

Results from the File format research project that are important to migration will be discussed in this document.

KB and migration

Appendix content e-Depot

In the appendix an example of the functional and technical characterisation of a scientific article from the e-Depot is given. Also, a first sketch of a functional characterisation of the complete content of the e-Depot is described.

2 Content e-Depot

2.1 Introduction

This chapter contains a list of archived objects in the e-Depot and future objects categorised on file format.

Based on the results of the File format research project obsolete formats or formats that are in danger of becoming obsolete are indicated.

The characterisation of the most common type of object, archived in the e-Depot, is described at the end of this chapter.

2.2 Archived objects in the e-Depot

The following file formats have been archived in the e-Depot (based on numbers from February 2006):

PDF 1.1	288.823
PDF 1.2	1.623.991
PDF 1.3	1.902.242
PDF 1.4	28.422
PDF 1.5	15.449
TIFF 5.0	4.163
TIFF 6.0	273.281
JPG	252.107 (attachments, not main files)

The DARE project will add a more heterogeneous supply of digital objects to the e-Depot:

- MS Office (Word documents)
- Photoshop documents.
- Etc.

The web archiving project will add websites to the content of the e-Depot:

- Websites can contain html, different kinds of image formats (JPEG, GIF) Flash, JavaScript and other forms of scripting, ASP, databases among others.

2.3 Characterisation of digital objects

To formulate migration projects and procedures, it is necessary to define what properties the original file has and which properties also need to be retained in the target file.

To characterise digital objects based on the file format is becoming more difficult now that many file formats are extended. A file format such as PDF is extended in functionality and can not be regarded as a standard “text document” when new functionality such as the incorporation of moving images in a PDF is used.

KB and migration

It is necessary to make a functional classification, next to a technical classification (based on file format). A functional classification is based on the idea of what function the digital object has for the public. A functional classification could have categories such as “publication from a scientific journal” or “informative website of government institution”.

At this moment, not many automatic options for characterisation exist. The software program Jhove generates technical metadata (MIX elements) such as which fonts are embedded or subset and what size the file is. This is characterisation on file level. It is necessary to make a classification at a higher level than on file level for migration procedures. The e-Depot contains too many objects to define a migration procedure per archived object.

Functional characterisation, concerning the 5 properties of a digital object (content, context, appearance, behaviour and structure) can be defined manually only.

The migration procedures for this research project have been based on migration of a publication in a scientific journal and its specific characterisation.

An example of characterisation of a scientific publication is presented in the appendix of this document. Publication from a scientific journal is the main category of objects that have been archived in the e-Depot.

A first draft of a functional classification has also been added to this document.

3 Possibilities and consequences of the use of migration at the KB

3.1 Introduction

This chapter discusses the possibilities and consequences of the use of migration at the KB, based among others on the preliminary results from the File format research project. The File format research project looks into which file formats, archived in the e-Depot, are in danger of becoming inaccessible and qualify for the use of migration.

3.2 Results File format research project

Research into the possible loss of accessibility of the publications archived in the e-Depot is still in progress (File format research project).
As yet no version of PDF or TIFF seems to be in danger of becoming inaccessible, although a random check test as part of the File format research project has not been performed yet (File format research project).

[Results File format research project]

3.3 Test Migration research project

For the migration research project it is still worth while to perform migration tests on digital objects archived in the e-Depot, even though these digital objects are not yet in danger of becoming inaccessible.

Based on the test and the test results the following items can be set up:

- Developments of test procedures
- Developments of migration procedures
- Drawing up of a list of tools needed for migration
- Drawing up of a list of vacant parts concerning migration processes that have not been tested
- Collection of information that can be used for the set up of a registry for preservation tools (a registry is part of the PLANETS project)

KB and migration

3.4 Application of migration at the KB

Options for the use of migration in the current situation:

- Normalisation of uncommon file formats

The DARE project is a concrete example of this and will receive a heterogeneous offer of file formats from the Dutch Universities. It is most likely that uncommon file formats will be delivered. Normalisation can be an option for the DARE project.

Consequences:

- Archiving the original version and a normalised version takes about twice as much space in the e-Depot.
- A normalised version also needs a digital preservation strategy. However, less effort has to be spent on the maintenance of a digital preservation strategy for a few standard formats than on a wide range of formats.

Future options for migration:

- Migration to a newer version of the same file format

Consequences:

- The National Archives of the Netherlands tested step by step migration to a newer version of the same file format. They compared this to migration in leaps to a newer version (some in-between versions are skipped). The results showed that a step by step migration generated more errors than migration in leaps.

- Migration on request

The CAMiLEON project and LOCKSS have performed a test with migration on request. Both projects have positive results: migration on request is feasible. If the conversion tool is designed well, it can partly be used for the long term.

- UVC

A UVC for images (JPG and GIF) has been developed. A UVC for PDF is too complex to develop within the migration research project.

I Appendix content e-Depot

KB and migration

I.1 Example of functional characterisation of an e-Depot article

The example article is a PDF version 1.3 file, displayed in Adobe Acrobat Reader 7.0.0. 14-12-2004

Article:

Predicting survival outcomes using subsets of significant genes in prognostic marker studies with microarrays

Shigeyuki Matsui^{1,2} ✉

¹Department of Pharmacoepidemiology, School of Public Health, Kyoto University, Yoshida Konoe-cho, Sakyo-ku, Kyoto 606-8501, Japan

²Translational Research Informatics Center, Foundation for Biomedical Research and Innovation, Minatojima-minami-machi, Chuo-ku, Kobe 650-0047, Japan

BMC Bioinformatics 2006, **7**:156 doi:10.1186/1471-2105-7-156

A blue **marking** indicates where potential “problems” might be encountered with reading and understanding the e-Depot article.

I have tried, while dividing the article over the five properties *content*, *structure*, *appearance*, *context* and *behaviour*, to discuss each property separately from each other. For example, for content the size of the font is not important. Structure defines a heading and subheading, but it is appearance that defines that a heading is in a larger font than a subheading.

It can not be recommended to characterise digital objects by file format only, now that many formats are extended functionally. A classification on type of file next to a classification on file format is useful. A ‘type of file’ means a classification such as text documents, still images, moving images and spreadsheets. An example follows:

Text documents Images * ...
 * Article
 * Manual
 *Documentation

or

Type of object	Functional purpose of object		
	Article	Manual	Documentation
Text documents			
Images			
Spreadsheets			

KB and migration

Content

Background

Genetic markers hold great promise for refining our ability to establish precise prognostic prediction for diseases. The development of comprehensive, gene expression microarray technology has allowed the selection of relevant marker genes from a large pool of candidate genes in early-phased, developmental prognostic marker studies for various cancers including diffuse large B-cell lymphoma [1], follicular lymphoma [2], acute myeloid leukemia [3], lung adenocarcinoma [4], and metastatic kidney cancer [5]. The selected genes will be further investigated in subsequent studies using technically simpler, but more reliable assays such as multiplexed quantitative reverse transcriptase polymerase chain reaction (RT-PCR) in formalin-fixed, paraffin-embedded tissue sections for routine clinical use [6,7]. Accordingly, the primary task in

Published: 20 March 2006

BMC Bioinformatics 2006, **7**:156 doi:10.1186/1471-2105-7-156

Received: 01 September 2005

Accepted: 20 March 2006

This article is available from: <http://www.biomedcentral.com/1471-2105/7/156>

© 2006 Matsui; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

BMC Bioinformatics 2006, **7**:156 <http://www.biomedcentral.com/1471-2105/7/156>

Page 2 of 9

(page number not for citation purposes)

early-phased prognostic marker studies with microarrays would be to select a small fraction of relevant genes for subsequent studies. To this end, multiple testing to identify genes associated with prognosis is typically adopted as primary analysis, which may provide a list of significant genes.

Prediction analysis using subsets of significant genes may supplement the primary analysis. It can provide information regarding predictive capability for subsets of significant genes. More importantly, provided that appropriate measures of predictive accuracy for survival outcomes are established, it may indicate another 'cut-off' for a list of significant genes on the basis of predictive accuracy through *gene filtering* other than the criteria to control false positives in multiple testing. Despite many methods for prediction analysis of survival outcomes proposed in bioinformatics and biostatistics literature, including application of partial least squares [8-10] and ridge regression

KB and migration

[11,12], most methods intend to use the full set of genes for prediction without regard to the primary analysis. In this article, we develop a methodology for predicting survival outcomes using subsets of significant genes. Key components in this methodology include building prediction models, assessing predictive performance of prediction models, and assessing significance of prediction results. Here, given the first two components, we can perform gene filtering. For each component, we consider particular specifications or procedures to illustrate the methodology. In building prediction models, we assume Cox proportional hazard models with a compound covariate [13,14]. In assessing performance of prediction models, measures of explained variation for Cox regression models [15-17] may not aim to measure the performance of prediction models on future patients, i.e., predictive accuracy. We propose to use the cross-validated log partial likelihood [18,19] to measure predictive accuracy. To assess significance of prediction results, we apply the permutation procedures in cross-validated prediction proposed by Radmacher et al. [14]. As an additional key component peculiar to prognostic prediction, we also consider incorporation of standard prognostic factors, because it is important to determine whether a new genetic marker adds prognostic information to that already contained in the more established prognostic factors [20]. The performance of the methodology will be evaluated using simulated data and real data from a lymphoma study.

KB and migration

Structure

Page 1

Left column

Heading paragraph: Background

Genetic markers hold great promise for refining our ability to establish precise prognostic prediction for diseases. The development of comprehensive, gene expression microarray technology has allowed the selection of relevant marker genes from a large pool of candidate genes in early-phased, developmental prognostic marker studies for various cancers including diffuse large B-cell lym-

Right column

phoma [1], follicular lymphoma [2], acute myeloid leukemia [3], lung adenocarcinoma [4], and metastatic kidney cancer [5]. The selected genes will be further investigated in subsequent studies using technically simpler, but more reliable assays such as multiplexed quantitative reversetranscriptase polymerase chain reaction (RT-PCR) in formalin-fixed, paraffin-embedded tissue sections for routine clinical use [6,7]. Accordingly, the primary task in

Page 2

Left column

BMC Bioinformatics 2006, **7**:156 <http://www.biomedcentral.com/1471-2105/7/156>

Page 2 of 9

(page number not for citation purposes)

early-phased prognostic marker studies with microarrays would be to select a small fraction of relevant genes for subsequent studies. To this end, multiple testing to identify genes associated with prognosis is typically adopted as primary analysis, which may provide a list of significant genes.

Prediction analysis using subsets of significant genes may supplement the primary analysis. It can provide information regarding predictive capability for subsets of significant genes. More importantly, provided that appropriate measures of predictive accuracy for survival outcomes are established, it may indicate another 'cut-off' for a list of significant genes on the basis of predictive accuracy

KB and migration

through *gene filtering* other than the criteria to control false positives in multiple testing. Despite many methods for prediction analysis of survival outcomes proposed in bioinformatics and biostatistics literature, including application of partial least squares [8-10] and ridge regression [11,12], most methods intend to use the full set of genes for prediction without regard to the primary analysis. In this article, we develop a methodology for predicting survival outcomes using subsets of significant genes. Key components in this methodology include building prediction models, assessing predictive performance of prediction models, and assessing significance of prediction results. Here, given the first two components, we can perform gene filtering. For each component, we consider particular specifications or procedures to illustrate the methodology. In building prediction models, we assume Cox proportional hazard models with a compound covariate [13,14]. In assessing performance of prediction models, measures of explained variation for Cox regression models [15-17] may not aim to measure the performance of prediction models on future patients, i.e., predictive accuracy. We propose to use the cross-validated log partial likelihood [18,19] to measure predictive accuracy. To assess significance of prediction results, we apply the permutation procedures in cross-validated prediction proposed by Radmacher et al. [14]. As an additional key component peculiar to prognostic prediction, we also consider incorporation of standard prognostic factors, because it is important to determine whether a new genetic marker adds prognostic information to that already contained in the more established prognostic factors [20]. The performance of the methodology will be evaluated using simulated data and real data from a lymphoma study.

KB and migration

Appearance

Methodology article

Open Access

Predicting survival outcomes using subsets of significant genes in prognostic marker studies with microarrays

Shigeyuki Matsui*^{1,2}

Address: ¹Department of Pharmacoepidemiology, School of Public Health, Kyoto University, Yoshida Konoe-cho, Sakyo-ku, Kyoto 606-8501, Japan and ²Translational Research Informatics Center, Foundation for Biomedical Research and Innovation, Minatojima-minami-machi, Chuo-ku, Kobe 650-0047, Japan

Email: Shigeyuki Matsui* - matsui@pbh.med.kyoto-u.ac.jp

* Corresponding author

Published: 20 March 2006

Received: 01 September 2005

BMC Bioinformatics 2006, 7:156 doi:10.1186/1471-2105-7-156

Accepted: 20 March 2006

This article is available from: <http://www.biomedcentral.com/1471-2105/7/156>

© 2006 Matsui; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: Genetic markers hold great promise for refining our ability to establish precise prognostic prediction for diseases. The development of comprehensive gene expression microarray technology has allowed the selection of relevant marker genes from a large pool of candidate genes in early-phased, developmental prognostic marker studies. The primary analytical task in such studies is to select a small fraction of relevant genes, typically from a list of significant genes, for further investigation in subsequent studies.

Results: We develop a methodology for predicting survival outcomes using subsets of significant genes in prognostic marker studies with microarrays. Key components in this methodology include building prediction models, assessing predictive performance of prediction models, and assessing significance of prediction results. As particular specifications, we assume Cox proportional hazard models with a compound covariate. For assessing predictive accuracy, we propose to use the cross-validated log partial likelihood. To assess significance of prediction results, we apply permutation procedures in cross-validated prediction. As an additional key component peculiar to prognostic prediction, we also consider incorporation of standard prognostic factors. The methodology is evaluated using both simulated and real data.

Conclusion: The developed methodology for prognostic prediction using a subset of significant genes can provide new insights based on predictive capability, possibly incorporating standard prognostic factors, in selecting a fraction of relevant genes for subsequent studies.

Background

Genetic markers hold great promise for refining our ability to establish precise prognostic prediction for diseases. The development of comprehensive, gene expression microarray technology has allowed the selection of relevant marker genes from a large pool of candidate genes in early-phased, developmental prognostic marker studies for various cancers including diffuse large B-cell lym-

phoma [1], follicular lymphoma [2], acute myeloid leukemia [3], lung adenocarcinoma [4], and metastatic kidney cancer [5]. The selected genes will be further investigated in subsequent studies using technically simpler, but more reliable assays such as multiplexed quantitative reverse-transcriptase polymerase chain reaction (RT-PCR) in formalin-fixed, paraffin-embedded tissue sections for routine clinical use [6,7]. Accordingly, the primary task in

early-phased prognostic marker studies with microarrays would be to select a small fraction of relevant genes for subsequent studies. To this end, multiple testing to identify genes associated with prognosis is typically adopted as primary analysis, which may provide a list of significant genes.

Prediction analysis using subsets of significant genes may supplement the primary analysis. It can provide information regarding predictive capability for subsets of significant genes. More importantly, provided that appropriate measures of predictive accuracy for survival outcomes are established, it may indicate another 'cut-off' for a list of significant genes on the basis of predictive accuracy through *gene filtering* other than the criteria to control false positives in multiple testing. Despite many methods for prediction analysis of survival outcomes proposed in bioinformatics and biostatistics literature, including application of partial least squares [8-10] and ridge regression [11,12], most methods intend to use the full set of genes for prediction without regard to the primary analysis.

In this article, we develop a methodology for predicting survival outcomes using subsets of significant genes. Key components in this methodology include building prediction models, assessing predictive performance of prediction models, and assessing significance of prediction results. Here, given the first two components, we can perform gene filtering. For each component, we consider particular specifications or procedures to illustrate the methodology. In building prediction models, we assume Cox proportional hazard models with a compound covariate [13,14]. In assessing performance of prediction models, measures of explained variation for Cox regression models [15-17] may not aim to measure the performance of prediction models on future patients, i.e., predictive accuracy. We propose to use the cross-validated log partial likelihood [18,19] to measure predictive accuracy. To assess significance of prediction results, we apply the permutation procedures in cross-validated prediction proposed by Radmacher et al. [14]. As an additional key component peculiar to prognostic prediction, we also consider incorporation of standard prognostic factors, because it is important to determine whether a new genetic marker adds prognostic information to that already contained in the more established prognostic factors [20]. The performance of the methodology will be evaluated using simulated data and real data from a lymphoma study.

Results

Gene filtering

The simplest approach of gene filtering is based on the marginal association between each gene expression and

survival time [1-5]. For patient i in the training set, let $h_i(t)$ be the hazard function and x_{ji} be the expression level for gene j . For gene j , we assume the univariate Cox regression model,

$$h_i(t) = h_{j,0}(t) \exp(\beta_j x_{ji}) \quad (1)$$

where $h_{j,0}(t)$ is the baseline hazard function and β_j is a parameter. Gene filtering is based on a test of hypothesis $\beta_j = 0$ (e.g., a score or Wald test [21]). Genes are typically ranked on the basis of the value of absolute standardized test statistic. Gene filtering can be based on the number of genes [4] or a P -value cut-off [1,2,5]. A standardized score or Wald test statistic for testing hypothesis $\beta_j = 0$ is asymptotically normal with unit variance and mean equal to $D^{1/2} \beta_j \sigma_j^{-2}$ where σ_j^2 is the variance of expression levels across patients for gene j and D is the expected number of events [22]. The gene filtering is thus based on the hazard ratio associated with a change of standard deviation in gene expression for a given number of events.

Prediction model

For the set of K selected genes (j_1, \dots, j_K), the compound covariate for patient i is defined as

$$c_i = \sum_{k=1}^K z_{j_k} x_{j_k,i} \quad (2)$$

where z_{j_k} is the standardized test statistic obtained in the gene filtering for the selected gene j_k ($k = 1, \dots, K$). The definition of the compound covariates weights by means of standardized test statistics has been suggested for generalized linear models in Radmacher et al. [14]. This weighting policy reflects the criterion in the gene filtering step. Another possible policy is to use an estimate of β_j in stead of z_j , as the weight for gene j (e.g., Beer et al. [4]). Our weighting policy gives higher weight to genes with larger variance, which would yield a more robust predictor for subsequent validation studies because the expression profiles for genes with larger variance would be more reproducible.

The compound covariate can be regarded as a prognostic index; patients with large values of the compound covariate may have poor prognosis. We assume the following Cox model to relate the compound covariate to the survival time,

$$h_i(t) = h_0(t) \exp(\psi c_i) \quad (3)$$

KB and migration

First page

Name journal: left alignment

Publisher logo: right alignment

Dividing line: line across width of the page

Sub title article: left alignment, font type GillSans-Light, font size 14 pt, font colour black

Open access image: right alignment, colour image

Title article: left alignment, font type GillSans-Bold, font size 16 pt, font colour black

Author: left alignment, font type Giovanni-Book, font size 15 pt, font colour black

And so on

KB and migration

Context

Title: Predicting survival outcomes using subsets of significant genes in prognostic marker studies with microarrays: methodology article
Author: Shigeyuki Matsui
Published in: BMC Bioinformatics
Year: 2006

This document is a scientific article.

KB and migration

Behaviour

1. Search in the document.
2. Navigation. Division between table of content (Quadrivalvular rheumatic heart disease, Introduction, References) en pages (2).
3. Print
4. Changing the document
5. Compose the document
6. Copying or extracting the textual content
7. Extract content for access
8. Making notes
9. Filling in of form fields
10. Signature
11. Making templates

Note: Where do the properties of the document end and do the properties of the rendering application begin?

In this example the properties of the document can be put “on” or “off” at creation time. Scrolling, for example, is considered a property of the application that renders the example document and not a property of the document itself.

KB and migration

1.2 Example technical characterisation of e-Depot article

Technical characterisation can be on file format level or file level. File level characterisation can exist of the output generated by Jhove. Information on file format level would be the presence of a file header that contains information like author or description for example.

Jhove output (excerpts) for this example article:

```
Jhove (Rel. 1.0, 2005-05-26)
Date: 2006-08-16 12:15:42 CEST
RepresentationInformation: test_migratie\t1_wordtopdf\1471-2105-7-156.pdf
ReportingModule: PDF-hul, Rel. 1.4 (2005-03-09)
LastModified: 2006-08-15 23:15:52 CEST
Size: 381971
Format: PDF
Version: 1.3
Status: Not well-formed
SignatureMatches:
  PDF-hul
ErrorMessage: Invalid destination object
Offset: 370422
MIMEtype: application/pdf
PDFMetadata:
  Objects: 477
  FreeObjects: 1
  IncrementalUpdates: 2
  DocumentCatalog:
    PageLayout: SinglePage
    PageMode: UseOutlines
  Outlines:
    Item:
      Title: Abstract
      Destination: edu.harvard.hul.ois.jhove.module.pdf.PdfSimpleObject@1ce2dd4
    Children:
      Item:
        Title: Background
        Destination: edu.harvard.hul.ois.jhove.module.pdf.PdfSimpleObject@122cdb6
      Item:
        Title: Results
        Destination: edu.harvard.hul.ois.jhove.module.pdf.PdfSimpleObject@1ef9157
      Item:
        Title: Conclusion
        Destination: edu.harvard.hul.ois.jhove.module.pdf.PdfSimpleObject@12f0999
    Item:
      Title: Background
      Destination: edu.harvard.hul.ois.jhove.module.pdf.PdfSimpleObject@11f2ee1
    Item:
      Title: Results
```

KB and migration

Destination: edu.harvard.hul.ois.jhove.module.pdf.PdfSimpleObject@3ecfff
Children:
Item:
Title: Gene filtering
Destination: edu.harvard.hul.ois.jhove.module.pdf.PdfSimpleObject@65a77f
Item:
Title: Prediction model
Destination: edu.harvard.hul.ois.jhove.module.pdf.PdfSimpleObject@1d7ad1c
Item:
Title: Predictive accuracy
Destination: edu.harvard.hul.ois.jhove.module.pdf.PdfSimpleObject@a61164
Item:
Title: Adjustment for prognostic factors
Destination: edu.harvard.hul.ois.jhove.module.pdf.PdfSimpleObject@bfc8e0
Item:
Title: Simulated data
Destination: edu.harvard.hul.ois.jhove.module.pdf.PdfSimpleObject@11d0a4f
Item:
Title: Lymphoma data
Destination: edu.harvard.hul.ois.jhove.module.pdf.PdfSimpleObject@18fd984
Item:
Title: Discussion
Destination: edu.harvard.hul.ois.jhove.module.pdf.PdfSimpleObject@111a775
Item:
Title: Conclusion
Destination: edu.harvard.hul.ois.jhove.module.pdf.PdfSimpleObject@91cee
Item:
Title: Acknowledgements
Destination: edu.harvard.hul.ois.jhove.module.pdf.PdfSimpleObject@4a63d8
Item:
Title: References
Destination: edu.harvard.hul.ois.jhove.module.pdf.PdfSimpleObject@1e0ff2f
Info:
Title: 1471-2105-7-156.fm
Author: csproduction
Creator: FrameMaker 7.0
Producer: Acrobat Distiller 5.0.5 (Windows)
CreationDate: Tue Aug 15 14:03:26 CEST 2006
ModDate: Tue Aug 15 14:10:44 CEST 2006
ID: 0x5c55fb76e1246f68adb2f0ee3c6e88d4, 0x7ef8d9594b611e50de0824ea629f539a
Filters:
FilterPipeline: FlateDecode
Images:
Image:
NisoImageMetadata:
MIMEType: application/pdf
CompressionScheme: Deflate
ImageWidth: 706
ImageLength: 706

KB and migration

Fonts:
Type0:
Font:
 BaseFont: KNLBEF+MT-Extra
 Encoding: Identity-H
 ToUnicode: true
Font:
 BaseFont: KNLANH+SymbolMT
 Encoding: Identity-H
 ToUnicode: true
Type1:
Font:
 BaseFont: Courier
 FirstChar: 32
 LastChar: 32
 FontDescriptor:
 FontName: Courier
 Flags: FixedPitch, Serif, Nonsymbolic
 FontBBox: -28, -250, 628, 805
 Encoding: WinAnsiEncoding
Font:
 BaseFont: KNLOJI+CXGUNB+Times-Italic
 FontSubset: true
 FirstChar: 80
 LastChar: 80
 FontDescriptor:
 FontName: KNLOJI+CXGUNB+Times-Italic
 Flags: Nonsymbolic, Italic
 FontBBox: 0, 0, 605, 653
 FontFile3: true
 Encoding: WinAnsiEncoding

```
XMP: <rdf:RDF xmlns:rdf='http://www.w3.org/1999/02/22-rdf-syntax-ns#'
xmlns:iX='http://ns.adobe.com/iX/1.0/'><rdf:Description about="
xmlns='http://ns.adobe.com/pdf/1.3/' xmlns:pdf='http://ns.adobe.com/pdf/1.3/'
pdf:CreationDate='2006-08-15T12:03:26Z' pdf:ModDate='2006-08-15T12:10:44Z'
pdf:Producer='Acrobat Distiller 5.0.5 (Windows)' pdf:Author='csproduction'
pdf:Creator='FrameMaker 7.0' pdf:Title='1471-2105-7-156.fm'/>
<rdf:Description about=" xmlns='http://ns.adobe.com/xap/1.0/'
xmlns:xap='http://ns.adobe.com/xap/1.0/' xap:CreateDate='2006-08-15T12:03:26Z'
xap:ModifyDate='2006-08-15T12:10:44Z' xap:Author='csproduction'
xap:MetadataDate='2006-08-15T12:10:44Z'><xap:Title><rdf:Alt><rdf:li xml:lang='x-
default'>1471-2105-7-156.fm</rdf:li></rdf:Alt></xap:Title></rdf:Description>
<rdf:Description about=" xmlns='http://purl.org/dc/elements/1.1/'
xmlns:dc='http://purl.org/dc/elements/1.1/' dc:creator='csproduction' dc:title='1471-2105-7-
156.fm'/>
</rdf:RDF>
```

KB and migration

1.3 Overview (preliminary) functional characterisation of e-Depot content

Below a first draft for a characterisation of the content of the e-Depot based on the archived digital objects is presented. In the table is described which of the five properties of a digital object should be retained minimally when migrations will be performed.¹

Type of object	Functional purpose of object			
	Publication	Educative program	Digital master cultural heritage	
Text document	Content Structure			
Still image			Content Structure Appearance	
Spreadsheet				
Interactive software		Content Structure Appearance Behaviour Context?		
Website	Content Structure Appearance? Behaviour?			

¹ The overview is a first draft for a functional classification. To supplement the principles of the e-Depot, an official classification of the content should be written, broadly supported by the departments of the KB.

KB and migration

	Context?			
--	----------	--	--	--