Project Number	IST-2006-033789
Project Title	Planets
Title of Deliverable	Report on usage models for libraries, archives and data centres: final results
Deliverable Number	D3
Contributing Sub-project and Work-package	PP/3
Deliverable Dissemination Level	External PU
Deliverable Nature	Report and model
Contractual Delivery Date	31st July 2009
Actual Delivery Date	
Author(s)	HATII – NANETH - SB

Abstract

This report outlines the results of the third (and final) iteration of a user study. The user study was targeted to identify user requirements for preservation of digital documents, records and data sets. User requirements are modelled in a user requirement model that can be used in a broader requirements model for digital preservation.

Keyword list

User studies; user requirements: usage model; preservation requirements

Contributors

Person	Role	Partner	Contribution
John W. Pattenden-Fail	author	HATII	
Laura Molloy	author	HATII	
Bart Ballaux	author	NANETH	
Annette Balle Sørensen	author	SB	
Filip Kruse	author	SB	
Jørn Thørgersen	author	SB	

Document Approval

Person	Role	Partner
Hans Hofman	Sub-project lead	NANETH
Frank Houtman	Reviewer	KB-NL

Distribution

Person	Role	Partner

Revision History

Issue	Author	Date	Description
0.1	John W. Pattenden-Fail	28/07/2009	Initial draft
0.2	Annete Balle Sörensen	29/07/2009	Revisions
0.3	Bart Ballaux	03/08/2009	Revisions

References

Ref.	Document	Date	Details and Version
PP/3-D1	Report on iteration 1	October 2007	
PP/3-D2	Report on iteration 2 and preliminary model	September 2008	
PP/6-D7	Completion of Automated Collection Profiling Service (2nd Iteration)	May 2009	

EXECUTIVE SUMMARY

This document reports about the third iteration of the user studies that have been conducted to identify user requirements for digitally preserved material.
The methodology that was used in the first two iterations of the study – a combination of data probes and contextual design – was dropped in favour of a more focused approach that would use actual materials the subject use in their everyday use. The three partners looked for distinct patterns of use based on their focuses (libraries, archives, and data centres), using the preliminary model (from the second iteration) as a starting point.
The statements from the interviews and workshops were analysed collaboratively by the researchers to refine the earlier model into a more simplified (yet also more detailed) final version. Aspects of the older model that were distinct to usage and context were removed; as an alternative, a series of guiding questions were created to adjust default priorities based on intended use.
This model (and the guideline questions) will be delivered to work package PP/4 for integration into the Preservation Planning Tool, PLATO.

TABLE OF CONTENTS

1. Introductory notes

This document reports on the final iteration of a study of usage requirements for preservation planning systems, specifically the PLATO software used in the Planets project. This is the third iteration of a study that began in 2007, looking at differences between users of libraries, archives, and data centres, – seeking common grounds for usage and isolating differences in behaviour for each environment.

In September 2008, the second iteration of this study was published (PP/3-D2), which determined a preliminary model of requirements, designed for import into the PLATO software. The purpose of this final iteration was to identify weaknesses in this preliminary model and publish a final, refined version that was created through further qualitative research.

The earlier iterations of PP/3 were in some ways introductory. The first iteration was primarily a test of the methodology and started exploring the basic categories of which usage could be perceived. The second iteration was the initial construction of a model, so similarities between the different users were organised into something with distinct measurements applied. The findings are reflected in the preliminary model itself, though there were some difficult to measure criteria that affected usage, such as the context of the information and the intent for which it was being used.

For this third and final iteration, the methodology employed in the previous two iterations was altered to respond to weaknesses identified by the earlier work. As the model is intended to be incorporated into production software, the issues relative to digital preservation were emphasised in the model. Aspects relating to context and intention of use were removed from the model itself, but a series of questions were developed in their place. These questions will aid in the selection of requirements and develop a preservation plan that reflects the context of the collection.

1.1 Role within Planets

The PP/3 usage model is designed for the PLATO software. PLATO utilises a large variety of requirements to weigh various preservation actions, applying a score to each. The usage model as presented to PLATO is designed to be incorporated into their codebase, though it is understood that the model must be fitted to the design of the software. The model here is thus somewhat conceptual and the PLATO developers can apply specific scientific units to each requirement as they work it into the technical framework of the program.

Therefore this usage model, while specifically designed for PLATO, can be seen a conceptual approach to defining user requirements for preservation planning. It does not formally fold back into any other component of the Planets project, but it can be referred to as a map of end-user perspectives.

2. Description of Methodology

2.1 Introduction

The initial design for the PP/3 study was to develop a qualitative methodology in the first two iterations and create a preliminary model by the end of the second; the third iteration was to use a quantitative approach (such as a questionnaire) to validate the preliminary model via a larger sample size.

After the publication of deliverable PP3-D2 in September 2008, the PP/3 work package reviewed the preliminary model and decided that the original approach of a quantitative questionnaire or

survey would not benefit the output goals of the work package. Though the preliminary output was a successful starting model of general user requirements, the following issues emerged after discussion and reflection:

- The requirements determined were general; more specific details would benefit the decision-making support system of PLATO. A quantitative follow-up would merely reinforce the generalisations of the requirements and fail to develop the model further.

- The three partners of the project (NANETH, HATII, and the SB) each focused on a specific user group (archives, data centres, and libraries). The preliminary model assimilated the different needs of each group into one model, making it pointless to look at these groups separately.

- A general methodology will result in a general model of requirements; to find more detailed information; a more specific approach to interviewing is needed.

- The probe method, which followed the *Contextual Design* approach to software development, was too time-consuming. It required a long-term commitment from study respondents, and the information gathered during the 5-week probe was not any richer than what could be determined from a brief interview. The long commitment (without any reparation) made it difficult to locate interested participants, as all researchers are busy.

- The level of enthusiasm from the participants in iteration 2 was uneven, resulting in inconsistent levels of detail. The lengthy commitment was most likely responsible for diminished enthusiasm.

- Asking participants about theoretical requirements for their information is fine, but the participants would provide more detailed requirements if they were working with actual examples of digital resources they were familiar with.

Given these weaknesses, it was agreed to revamp the final iteration of PP/3 with the goal of producing the best possible model. The differences in user groups would be explored, not brought together; it was decided that further qualitative research would be the route for achieving this instead of a quantitative approach.

Each partner in the study again chose users from their focus; as before, NANETH selected users of archives, HATII selected users of data centres, and the SB selected users of libraries. The user was the centre of the study, and each partner tailored their approach specifically to what was appropriate for each participant. The preliminary model (output of PP3-D2) was the starting point for questioning, as well as a framework for the analysis. The ‘affinity analysis’ method utilised in the first two iterations of the study was abandoned, as we were no longer seeking affinities between the three user groups. We were seeking more specific clarification of the existing requirements as well as looking for new requirements, and looking for differences in usage between the three target areas.

2.2 Users of libraries (SB)

The Statsbiblioteket in Århus, Denmark, approached their group of library users by selecting researchers working in several areas connected to the SB. Three collections were targeted: The national collection of newspapers, represented by different paper and digital versions of Politiken (a national newspaper), the Søren Kierkegaard Collection (the digitalisation of the collected works of Søren Kierkegaard), and the WebArchive (an archive of Danish web-pages). Scenarios were presented in a group workshop on June 11, 2009 (attended by three researchers from Department of History and Area Studies (Aarhus University), Department of Information and Media Studies (Aarhus University), and Institute of Literature, Culture, and Media Studies (University of Southern Denmark), respectively. Group discussion ensued regarding the differences in presentation between original content and potentially migrated or altered content.

The workshop involved group participation and brainstorming, and each collection yielded specific responses. The requirements vaguely described in the preliminary model (PP/3-D2), such as “look and feel”, were explored more thoroughly as users shared their experiences both actual and

potential. Likewise, definitions of copy versus original very generally discussed. Different priorities emerged from each archive, and the context in which material might be used became emphasised.

2.3 Users of archives (NANETH)

The records that were used in this participatory user research were prints of electronic records. One set of experiment records consisted of three various representations of an original Word Perfect 5.1 record and one set consisted of two representations of e-mails. In addition to these representations, some different representation methods for metadata were shown, primarily prompting the user to indicate which metadata were necessary and meaningful to him.

For the Word Perfect 5.1 record, one representation was made using Dioscuri, the modular emulator of the National Archives and the National Library of the Netherlands. It represented a record that was very similar to the “original” (look and feel). A second representation was a normalised version, using XENA, the normalisation tool of the National Archives of Australia, lacking a lot of the original lay-out and including some small mistakes. A third version was shown in QuickView Plus, a viewer that is capable of representing Word Perfect 5.1 files, and giving a representation that mostly preserves the layout, but also includes some mistakes and misinterpretations.

In the case of the e-mails in .mpg format, one representation was shown in XENA and two representations were shown in QuickView Plus.

Questions during the one to one and a half-hour long session focused on the preferences of the two users for a specific representation, the reasons for this, preferences for availability of metadata, etc.
The participant was asked to think aloud, and to indicate what he thought of the various representations. Based on the different sets of material, it was possible to get more detailed insights in some of the issues that were raised during iteration 2.

2.4 Users of data centres (HATII)

HATII again targeted users of data centres, this time looking to split between humanities and science-based data centres. One participant worked regularly with a biological database of protein strings, and frequently wrote documentation for the database as well as scripts and source code to manipulate the data. The other was a user services manager for an archaeology data centre in the UK who is responsible for resource management, use, promotion, and improvement.

An initial, introductory interview was conducted where the participants provided background on the data centre, and how they worked with it. This replaced the 5-week probe period, as the participant provided a list of frequent applications and digital resources used on a regular basis in their work up-front, instead of in daily diary form. The participants then provided HATII with actual sample documents that they use, representing a cross-section of their daily work. HATII then manipulated these samples, creating a series of variations to emulate potential migrations or alterations that may occur during preservation actions. Close attention was paid to the preliminary model of user requirements, as migrations were made that intentionally altered aspects relating to requirements that the interviewers were hoping to elaborate on, such as “design and look” or metadata. Some “bad” migrations were intentionally included to produce more stimulating discussion, for example a wiki was converted from HTML to PDF using an amateurish online tool, which resulted in clipped image boundaries and overlapping, unreadable text.

After these examples were ready, the participants and interviewers met again. Here, the participants were presented with potential scenarios of preservation effects. The original documents were compared side-by-side with their altered version, - sometimes against several alterations at once. A discussion ensued where the participant evaluated the different migrations with respect the original and identified reasons in which some migrations would be preferable to others. Their comments provided a basis for which a more refined, final model could be constructed.

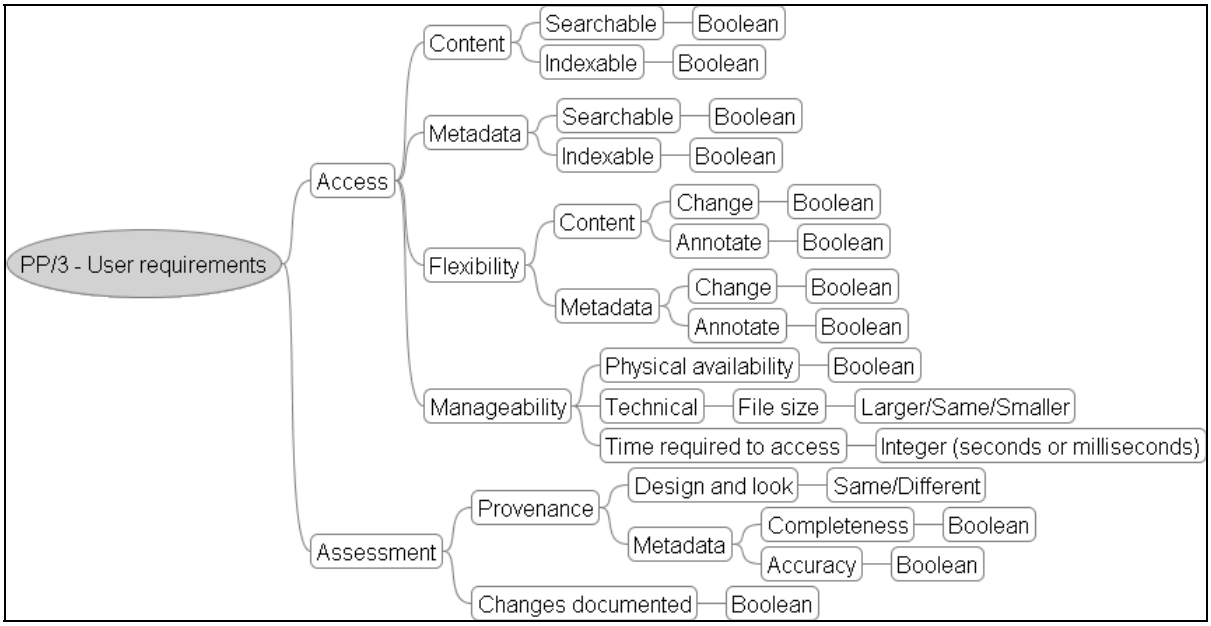
3. Analysis of results

The analysis of these interviews and workshops was conducted in Copenhagen, Denmark on 24 June, 2009, at Det Kongelige Bibliotek (The Royal Library). Abandoning the affinity analysis method, the researchers went through each partner's interview and workshop transcripts in sequence, looking for new requirements. Then the preliminary model from PP3-D2 (represented as a FreeMind mind-map) was worked through, node-by-node, looking at how each of the earlier requirements could now be refined.

3.1 Changes from the preliminary model

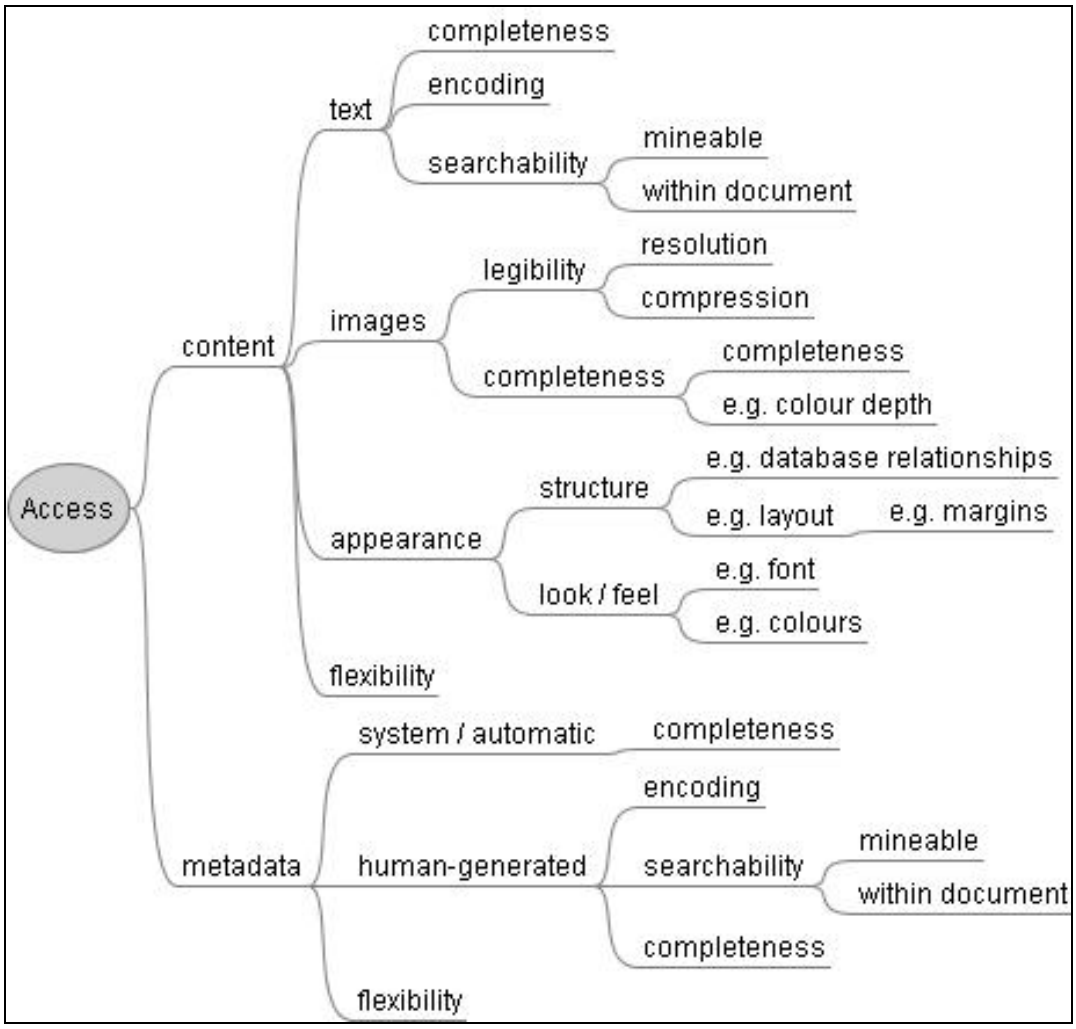
The analysis led to the creation of the "final" model (Fig. 2), which greatly simplifies the preliminary model (Fig. 1) (as described in PP3-D2).

Fig. 1. Preliminary model (from PP3-D2)



The preliminary model (Fig. 1) attempted to deal with issues of context and use through its two initial nodes, 'Access' and 'Assessment'. Through the iteration 3 interviews and workshop, it was decided that the requirements for 'Assessment' exist outside of the realm of preservation planning requirements, and therefore do not belong in the model. Some of the requirements in the final version (Fig. 2) were moved to a different place in the model, while others were eliminated.

Fig. 2. Final version of model



The final model thus is concerned only with access, and the root node is labelled as such (although there is no real need to label the root as anything). Repeating themes of content and metadata were seen as confusing, and furthermore, all requirements could be described as falling under one or the other. Thus, the model was reorganised to place content and metadata as the two initial branches, under which all requirements will be classified.

The specifics of each requirement will be described below in Section 4, but there are a few new aspects of the model to explain here. The previous requirement heading of “Content” was refined into specific areas, such as text, images and overall appearance. Appearance characteristic such as font, margins, and layout were placed here, and removed from the now-defunct “Assessment” section. The participants in the study provided information that allowed the final model to include the new requirements of database relationships, specific aspects of images, and character encoding. The area of “Manageability” was broken up, with most of those requirements falling now under the umbrella of “system/automatic” metadata – file sizes, for example, which affect the manageability of information.

3.2 Usage across collections

The goal of determining user requirements for three different target groups poses the problem to properly reflect the needs of each audience, there could end up being three different models. The nature of designing a model to be used in decision support for preservation planning must allow for all possible desires, so the problem is how to approach the differing models inclusively.

The obvious idea would be to propose three distinct models: one model of usage requirements for libraries, one for archives, and one for data centres. However, the models would be essentially identical; most of the actual requirements would exist across all three models, just prioritised differently. Moreover, upon closer inspection, some requirements would be more important under certain usage scenarios that would depend on the context and purpose of how the material is to be used, as opposed to merely the type of collection being dealt with.

Though the “assessment” section was removed from the model, it still was an integral part of the user experience, creating a problem on how to represent assessment behaviours in the model. Many of the old assessment requirements could affect the access requirements. For example, if the provenance of a document was of high importance, then the appearance of the document (such as fonts, colours, etc.) would be more important than for a scenario in which provenance was not in question. These eliminated requirements can be very important to the preservation planning process, even if they cannot be directly represented as quantifiable attributes used to grade migration or emulation processes.

Furthermore, users from the different types of collections that were studied (libraries, archives, and data centres) may utilise information differently because of their environment. For example, a scientist using a data centre of astronomical information may have no concern for appearance, since the data may just be tables of numbers. Yet an archival user, who is often dealing with scanned or digitised information, may be very concerned with the completeness or resolution of a scanned image (in one example, a user was looking for handwritten notes in the margins of scanned archival documents that were not in OCRd adaptations). An archivist or preservation officer at either of these organisations would take these intents into consideration when selecting requirements for preservation planning.

Therefore, we propose a series of questions to be asked during the preservation planning process that, depending on the answers, will alter the priorities of the requirements. These questions do not break down strictly into the roles of libraries, archives and data centres, as those boundaries are somewhat artificial; any type of usage can certainly occur in any type of environment. However, the feedback from the interviews in this iteration indicated several such scenarios, which were discovered through probing the users about their own work within the context of their environment.

These are six simple questions that can be asked to *pre-weigh* some of the requirements with values that would be more appropriate for the usage intended:

1. Is the content digital-born?
2. Is the content likely to be represented in a paper/analogue format?
3. Is the appearance of this content relevant?
4. Do you want this content to be searchable?
5. Do you want to alter/edit a personal copy of this content?
6. Do you want to be able to check the provenance of this content?

It was decided that these guidelines will weigh the requirements in an abstract manner, and the PLATO development team, should they choose to incorporate the guidelines into the application, can determine specific values for the requirements. Many requirements will have a specific numerical scale, such as describing the resolution of an image or the size of a file. Others are merely defined in relation to the original documents, such as “less complete” or “lower resolution”.

Rather than assign a specific value scale to each requirement (as was attempted in the preliminary model), the final model leaves this somewhat more abstract. The PLATO development team is free to adapt these requirements into their system as they see fit. Indeed, some requirements may already exist in other areas not covered by the PP/3 work package.

The answers to each of the guideline questions are intended to alter the weight of a specific group of requirements. The rankings of the requirements (ultimately represented numerically in PLATO) produce scores for different preservation actions, such as migrations. These guideline questions will either increase or decrease the importance of these requirements before the preservation actions are scored. After each question, each requirement will be affected in one of three ways, depending on the answer:

1. No change (the question is not applicable to this requirement)

2. Lower the priority (the requirement is less important because of the answer to the question, so its existing value can be lowered)
3. Increase in priority (the requirement is somewhat more important because of the answer to the question, so the existing value can be raised somewhat)

Below are descriptions of each question with some example scenarios, and the outcomes that should result from a yes or no answer. Detailed definitions of each requirement are available below in section 4 of this document. If any questions cannot be answered clearly (due to uncertainty, or both a "yes" and "no" answer applying at the same time), they should be skipped with no alteration to the existing requirements.

3.2.1 Is the content digital-born?

If answered "yes":

Lower: Content:Images:Legibility, Content:Images:Completeness

Increase: Metadata:Automatic:Completeness, Metadata:Human-generated:Completeness.

If the answer is "no"

Increase: Content:Images:Legibility, Content:Images:Completeness, Content:Appearance:Layout, Content:Appearance:Look/feel

No other requirements are altered beyond normal.

If the content is not digital-born – say, if it is a scanned version of a paper document – then a higher priority must be assigned to the legibility and completeness of images, as the digitisation process may introduce unwanted artefacts. A user who is working with digitised images may demand a higher accuracy of representation than if the images were digitally synthesised. Likewise, the priorities can be lowered for digital-born data.

"Automatic" metadata, which includes values like the timestamp of the file, internal attributes native to a file format or metadata generated by the information system in which the file resided, are of somewhat more importance in digital-born data. The same goes for human-generated metadata such as notes and comments, abstracts, tags, etc.

In the case that a preservation plan is being created for a collection that mixes digital-born and digitised content, this question can be ignored and no priorities will be adjusted.

3.2.2 Is this content likely to be represented in paper/analogue form?

If answered "yes":

Lower: Content:Appearance:Structure

Increase: Content:Appearance:Layout, Content:Appearance:Look/feel

If answered "no":

Lower: Content:Apperance:Layout, Content:Appearance:Look/feel

Increase: Content:Appearance:Structure

This question is not as straightforward as it seems. While generally one would answer "yes" to data from libraries or archives, one would almost certainly answer 'no' for scientific data such as databases, source code, or raw scientific data. The Content:Appearance:Structure requirement refers to the internal structure of such data like the relationships between tables in a database, or

whitespace and margins in source code. A migration must score high in preserving the relationships between these tables, or it will ruin the integrity of the data.

Likewise, a written report is likely to be represented in paper form, so a moderate importance is assigned to aspects of layout and look/feel. This question is not about provenance, where one might require a very strict adherence to original fonts and design elements. Instead, this is a requirement for general use; users indicated that a minimal adherence to layout and design is helpful in content that could be potentially printed.

In the case of a digitised magazine article or handwritten archival record, one would answer 'yes' here, which would raise the priorities of the layout and look and feel somewhat; if one answers "yes" to question #6 (regarding provenance), these requirements will be increased further.

3.2.3 Is the appearance of this content relevant?

If answered "yes":

Increase: Content:Text:Completeness, Content:Text:Encoding, Content:Images:Legibility, Content:Images:Completeness, Content:Appearance:Structure, Content:Appearance:Layout, Content:Appearance:Look/feel

If answered "no":

Lower: Content:Appearance:Structure, Content:Appearance:Layout, Content:Appearance:Look/feel

This question is intentionally left fairly interpretable. Note that 'appearance' does not merely mean a paper/analogue appearance; it may mean the appearance of the content on the screen, or the formatting of whitespace within a Python script. All visual requirements are emphasised at high priority if the question is answered "yes".

If answered "no", i.e. if the appearance does not matter much, and just the raw data or content inside is important, then we can lower the importance of these requirements.

3.2.4 Do you want this content to be searchable?

If answered "yes":

Increase: Content:Text:Completeness, Content:Text:Encoding, Content:Text:Searchability, Metadata:Automatic:Completeness, Metadata:Human-Generated:Completeness, Metadata:Human-Generated:Encoding; Metadata:Human-Generated:Searchability

If answered "no", no requirements are affected.

Many of the discussions in these user studies revolved around the subject of searching. The ability to search through information is a high priority in many instances, though there may also be situations in which it is not so important. If the content is to be searchable, then formats that allow searching are clearly to be prioritised; to follow, metadata (which is often searched by indexing tools and search engines) and other aspects of content are highly important.

3.2.5 Do you want to alter or edit a personal copy of this content?

If answered "yes":

Increase: Content:Text:Completeness, Content:Text:Encoding, Content:Images:Completeness, Content:Flexibility, Metadata:Flexibility

If answered “no”:

Lower: Content:Flexibility, Metadata:Flexibility

In these user studies, requirements changed depending on what users intended to do with the information. In our sample scenarios, the users sometimes wished to use the preserved information in their own work. In other situations, however, they would only need to refer to it, the “read-only” situation, as it was called in our analysis. The requirements labelled here as “flexibility” describe actions where a user will edit or re-use aspects of the information and are thus highly important if a user intends to work with the content this way. Obviously, if the information is only being used as a reference source, then it is not necessary to prioritise the ability to alter or annotate content.

3.2.6 Do you want to be able to check the provenance of this content?

If answered “yes”:

Increase: Content:Images:Completeness, Content:Appearance:Layout, Content:Appearance:Look/feel, Metadata:Automatic:Completeness, Metadata:Human-Generated:Completeness, Metadata:Human-generated:Encoding, Metadata:Human-generated:Searchability

If answered “no”:

Lower: Content:Appearance:Layout, Content:Appearance:Look/feel

This final question addresses a major arm of our preliminary model, which was labelled as “Assessment”. Users often refer to visual/design elements to provide the authoritative basis for a piece of information. For example, layout, page headings, fonts and logos may be the visual proof, to a user, that a digital version of a journal article did indeed come from that journal. Therefore, if a user indicates that provenance is important to them, then it is important to preserve all images and visual aspects of content, a well as metadata that may also indicate the provenance of information more explicitly.

3.3 Collection profiling

These requirements will be applied to a collection undergoing the preservation planning process, so the nature of the collection will likely affect the priorities of the specific preservation plan. The interviews and workshops did not collect specific requirements for building collection profiles, but the statements of participants informed the research staff in developing a conceptual framework (in conjunction with the PP/6 work package).

The PP/6 work package of the Planets project has recently completed the second iteration of its collection profiling work, publishing the report ‘Completion of Automated Collection Profiling Service (2nd Iteration)”. This report outlines the development and use of the DROID tool, an automated tool for building a collection profile, with the goal of automating the preservation planning process.

A collection profile should contain two parts: identification information, and event history information. A description of these elements can be constructed through analysis of DROID’s technical characteristics and the structural elements of collections referred to by users in these field studies.

3.3.1 Identification Information

Identification information is what distinguishes a collection – the properties that define and describe the collection. The identification information of a collection can be broken down into two components: technical information and intellectual information.

3.3.1.1 Technical identification information

Technical information can usually be automatically collected by a tool such as DROID. Information such as the size and format of a file, as well as system-level meta-data such as modification times and ownership/permissions all falls under the domain of technical information. The DROID tool uses signatures to identify files in a collection, which could be used to group or otherwise categorise files for automatically triggered preservation actions.

3.3.1.2 Intellectual identification information

Intellectual information in a collection profile is a somewhat more conceptual. This information could include elements related to groupings (based upon some meaningful criteria) such as the number of files in a group or a hierarchy of information based upon content or other factors. The provenance of the collection can be established here, as well as time-based information identifying the content as belonging to a range or milieu.

Intellectual identification aspects may be more difficult to automatically extract using a tool such as DROID due to the often human-defined criteria for organisation and classification. End-users who were interviewed for the usage model studies indicated that they often used contextual elements for identification, such as which other documents might co-exist in a folder. These groupings may lead to differing requirements, based on the context/role of the information.

Intellectual identification information will differ depending on the institution and field.

3.3.2 Event History Information

The event history information of a collection profile characterises the transformations that a collection undergoes. File migrations, data refreshment, and other major changes can be profiled with any accompanying metadata such as date ranges.

Some of this information will be easy to automatically extract with a tool such as DROID, while other information will be dependent on the quality of the metadata delivered within the collection.

3.3.3 Collection profiling in use

The ability to trigger preservation actions via a collection profiling tool is enticing and thus raises the issue of how user requirements could also be automated. The guideline questions as described in section 3.2 could be somewhat automated, perhaps in conjunction with a DROID output report, but only if aspects of the collection were broken down into smaller components.

Most collections containing mixed information will render all of the guideline questions inapplicable, since they would likely be answered both “yes” and “no” at the same time. In future development, it may be worthwhile to apply each set of questions to smaller subsets of a collection, grouped among similar properties such as “all source code” or “all written Word documents”. Just as the guideline questions attempt to pre-weigh requirements specifications based upon use, some sort of layer between DROID and PLATO could attempt to pre-answer some guideline questions, or even eliminate them, by providing standard pre-ranked requirements for information types. This runs the risk of assuming what a user’s usage intentions are, which would defeat the purpose, so any requirement selection based upon collection must only be offered as “recommendations”.

4. Description of the model

The final model (Fig. 2) is the culmination of three iterations of user studies. This version is significantly simpler than the preliminary model developed during the second iteration, due to the separation of assessment/context requirements into the guideline questions described in section 3.2.

Below are descriptions of each node on the requirements model. These descriptions are intended to clarify the terminology we have chosen, and if necessary, provide some examples of how the requirement might be used. More potentially specific subrequirements are mentioned though not

explicitly indicated in the model; we leave this to the discretion of the PLATO implementation team, particularly as some of these subrequirements may exist already in PLATO.

4.1 Content

The Content heading includes all text, images, sound, video, or otherwise “main” content of a digital resource. Most interaction with the user is conducted through the content branch.

Content >Text

Requirements under this heading refer to textual elements of a digital resource.

Content > Text > Completeness

The completeness of text describes how much of the text remains after a given migration. Some aspects of text may be lost when converting from one format to another, for example accented characters. In visually-oriented documents such as scanned or OCRd data, parts of the text may end up clipped due to alterations in margins or other areas.

Completeness could be measured in some sort of numeric score, though it may be difficult to standardise how complete a text is through a ranking. Most likely, the requirement will be measured as a Boolean score, i.e. the text is either complete, or not. Preservation action scores can document what elements will be lost, so perhaps a scale could be constructed where certain aspects of the loss are selectable. This would not be a linear/numeric score but rather “multiple choice”. However, it would be unlikely that any preservation plan would ever NOT require the text to be complete.

One user in the study, when faced with a migration that eliminated captions from some images and clipped a few characters from the end of each line (due to an incorrectly calculated margin) stated that he would be happy with the slight loss of text because the essence of the information was maintained and he could easily understand the document despite the missing elements. This standard for use, however, may be a very small minority.

Content > Text > Encoding

The various encoding standards for text (such as UTF-8, Latin1, Chinese, etc.) can create major problems if the encoding is not preserved. For example, a Scandinavian language that uses letters such as ö, ø or å may be rendered incomprehensible if the encoding is not correctly preserved. If this requirement were set with a high priority, it would be essential that a preservation action maintain the correct encoding. This can also be a problem with text inside of databases or other content that is not a traditional “document”. Older formats may not be as compatible with modern encodings such as UTF-8; migrations may results in incorrect translations and ambiguous or missing characters.

Content > Text > Searchability

The ability to search for text within a document is a property of the document’s file format. For example, a PDF-A1a compliant file allows for searching within the text, but a JPG of a scanned page of text is not searchable. A user who has a high priority for searchability will require a preservation action to result in a searchable file.

Though this requirement refers to the ability to search within the document itself, it is also related to searching among a larger class of documents. If a document is searchable, it most likely allows for automatic extraction of its text via data mining software or search engine robots. Many of the interviews from this iteration and the previous one indicated that the use of search engines is very important in everyday work. Content > Text > Searchability > Mineability could potentially be a subrequirement here, though it may not be necessary if all searchable formats are also mineable.

Content > Images

Requirements under this heading refer to image information in digital content. This may be the entirety of the content if the content is an image file like JPG or PNG, or it may be just the image elements of a Word document or other mixed format. Video content can be grouped here as the Content > Images requirements should apply.

Content > Images > Legibility

Images are easily altered through the various algorithms and compression schemes used to represent them digitally. The Legibility requirement ensures that the user can comprehend the image to their specified level (as compared to the original).

From our interviews, we have found two sub-requirements for Content > Images > Legibility, both quantifiable: *i*) Content > Images > Legibility > Resolution measures the resolution of a still or motion image; a migration could affect the resolution through downsampling or resizing, and thus the legibility of the content could be compromised for the user. *ii*) Content > Images > Legibility > Compression is a measure of to what level compressed image formats employ the compression; higher levels introduce artefacts into the image that also compromise legibility. The requirement could be chosen as a comparative value against the original, i.e. the migration compression level must be equal or lower than the compression level of the original. Alternatively, a numeric value could be specified as a maximum allowable level of compression, similar to the quality scale that can be set when saving a JPG or other compressed image from Adobe Photoshop or similar software.

Content > Images > Completeness

Like the textual completeness requirement, image completeness refers to the availability of the entire image. Some preservation actions may clip or alter the images to fit within bounding boxes or other document constraints. By requiring that images must be complete, only migrations that are proven to have non-destructive image conversions will be allowed.

As a sub-requirement, colour depth is an aspect of an image that refers to its completeness; while the full area and resolution of a 24-bit colour image may not be altered, depth of field in colour may be lessened and thus the image's value will be compromised (Think of a full-colour image rendered in black and white, or 4-colour).

Content > Appearance

Appearance requirements don't apply to the actual content, but to the presentation of it. Interviews in this iteration and the previous one brought forth a wide range of opinions regarding the value of potentially non-essential attributes: font face, document margins, and background logos/watermarks (to name a few). We have subdivided Appearance requirements into two categories: Content > Appearance > Structure and Content > Appearance > Look/feel.

Content > Appearance > Structure

The Structure requirements refer to elements that define the structure of content. In a visual sense, the structure may define layout parameters such as margins and paragraph spacing. Thus, Content > Appearance > Structure > Layout can serve as a category for these attributes, with sub-requirements defined for specific aspects of the layout (margins, paper sizes, columns, etc.).

The preservation of complete databases, an issue that emerged from the interviews and discussions, applies to the Structure requirement. A database, being a series of tables with relations and indexes, relies upon the structure of these aspects for its integrity. Thus, a preservation plan that included complete databases would emphasise the structure requirement.

Content > Appearance > Look/feel

The term “look and feel” collects aspects of visual style present in digital content. On a most fundamental level, this would be the fonts and colours of a document, or the stylesheet aspect of web content. Content > Appearance > Look/feel can be further subdivided into quantifiable requirements for these aspects: fonts, colours, headings and logos, watermarks, etc.

Look and feel often plays a strong role in assessing the authenticity of a document. Users have indicated that often use design and layout elements to validate the provenance of digital content: For example, text from a journal will be considered authentic if the look and feel from the journal’s analogue form is reproduced. Of course, this varies among different users and situations, but appearance is often required depending on usage situations and context.

Content > Flexibility

The Flexibility requirement refers to the ability to modify a document. This will be a property of the format that the content is migrated to; for example a Word document allows one to add comments and track changes, while some PDF formats are read-only.

In the preliminary model, the Flexibility requirements were defined as “Changeable” and “Annotatable”. These requirements can remain as subrequirements of Content > Flexibility.

4.2 Metadata

The metadata requirements have been separated from the content branch and subdivided into two types of metadata. Creators have indicated that they understand the value of metadata for preservation and retrieval purposes, but the time constraints involved in the administration of their research often prevent them from properly completing metadata fields in the various applications they use. One person offered the opinion that adding metadata for preservation purposes should be mostly voluntary, stating, “You should really have to think carefully about what tasks are mandatory. And leave plenty of fields for people to add data voluntarily.” Thus we have separated metadata into two categories: automatic and human-generated.

Metadata > Automatic

“Automatic” metadata refers to information that is not user-generated. For example, all files will have a timestamp set by the operating system, and in some cases users would require this timestamp to be preserved (In one interview, the respondent indicated that he would use file dates and times as a type of verification, when trying to distinguish between different examples of similar files).

Additionally, automatic metadata might come from the software that creates the file, and not just the operating system. Adobe Photoshop and Microsoft Word embed metadata automatically into their formats, containing information about the creator and modification times. Migrating to another format might eliminate this metadata, which could contain valuable information for the end user.

Most digital cameras embed metadata with information about the hardware of the camera into the images. This is easily lost in image conversions but may be of importance to researchers.

Finally, systems in which documents reside, like electronic records management systems or digital repository systems, generate their own metadata (about e.g. management and preservation, but also about use, etc.) that are linked to the documents. These metadata can have value for users when assessing provenance or adequacy of the quality of a document.

Metadata > Human-generated

Metadata that must be added by hand falls under the Metadata > Human-generated requirements. This would include information like keywords, abstracts, and comments added to documents that are not part of the main content, and not generated automatically by the software.

This type of metadata, because it relies on the efforts of its creators, will vary in consistency.

Metadata > Human-generated > Completeness

Like textual or image content, the completeness of human-generated metadata can be easily measured when compared between “before” and “after” versions of migrated content. A simple check for the presence of the metadata will allow a ranking to be applied to various preservation actions. It is also important to determine if the entirety of the metadata is translated. For example, if a Microsoft Word document has a long comment added into its properties field (found by the File->Properties menu), it would be important to check if the comment has not been shortened or truncated by a migration to another format.

Metadata > Human-generated > Encoding

Like the Content > Text > Encoding requirement, human-generated metadata (because it is often text) can have similar encoding issues. This requirement will just verify that the encoding is preserved with metadata as well as content.

Metadata > Human-generated > Searchability

One of the most important requirements for usage is the ability to search through metadata, such as keywords. This aids in the retrieval of information from large repositories and allows users to find files even among their own collections. Like the Content > Text > Searchability requirement, this requirement ensures that metadata can be searched (and ideally, mined and indexed by search engines).

Metadata > Flexibility

The ability to modify metadata – either editing existing metadata or adding new values – is important in many usage scenarios. A preservation action should measure a migrated format’s value for metadata flexibility if this is required by a preservation plan. Annotations and alterations, like in the Content > Flexibility requirement, could be defined as specific subrequirements. It may even be the case that the original format of a document does not allow flexibility with metadata, but the migrated format can add this as a feature.

5. Conclusions

No study like this can ever be truly complete. Though digital preservation is in many ways about attempting to develop a “future-proof” system, emerging technologies will inevitably produce new demands. This model attempts to address the nature of user requirements theoretically, and provides an open framework for future technologies.

Through three iterations of our research, we have found an approach to mapping user priorities into practical requirements through a model, with the addition of the guidelines questions to better emphasise the needs of the user. The PLATO development team will hopefully implement this theoretical model in a practical manner and enrich the preservation planning tool.

Although the preliminary assumption was that users of libraries, data centres and archives would come up with various requirements, user feedback suggested that not the type of user, but the type of usage that determines the requirements. By combining six questions with a tree of abstract requirements, a weighting of the requirements is reached for the implementation in a decision support tool.