Project Number	IST-2006-033789
Project Title	Planets
Title of Deliverable	Report on usage models for libraries, archives and data centres, results of the second iteration
Deliverable Number	D2
Contributing Sub-project and Work-package	PP/3
Deliverable Dissemination Level	External
Deliverable Nature	Report and model
Contractual Delivery Date	31 8 2008
Actual Delivery Date	29 8 2008
Author(s)	HATII – NANETH - SB

Abstract

This report outlines the results of the second 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 Pattenden-Fail	author	HATII	
Bart Ballaux	author	NANETH	
Annette Balle Sørensen	author	SB	
Filip Kruse	author	SB	
Jørn Thøgersen	author	SB	

EXECUTIVE SUMMARY

This document reports about the second iteration of the user studies that have been conducted to identify user requirements for digitally preserved material.
The methodology that was introduced in the first iteration of the user study –a combination of data probes and contextual design– was slightly adapted to be used in this iteration. Each of the three partners approached users who regularly make use of archival material, library material, or data sets. The users described their research methods over approximately four weeks, and were interviewed at the end of the probe period.
On the basis of their feedback in the diary/data probe and interviews, statements of user opinions were gathered and clustered in an affinity analysis. This collaborative analysis revealed the themes that are most important to users in archives, libraries and data centres.
These themes are the backbone of the model that is presented in this report. Originally beginning as a model focused on three phases of research (search, assess and use), it was transformed into a model that is more object-oriented.
This model will be delivered to workpackage PP/4 for integration in the Preservation Planning Tool.

TABLE OF CONTENTS

1. Introductory notes

This report includes results of a user study in which library, archival and data centre users were asked to identify requirements that matter for their research. The methodology to interact with users was constructed with the intent not to influence the input of users in one or another direction. As a result, some preliminary notes have to be emphasised.

1. Users were asked to indicate all issues they were confronted with during the research process. As a result the requirements that were identified are not always relevant for digital preservation as such. Some of them are for instance about retrieval of information or about the availability of a paper copy. The fact that users emphasise other issues than those related to digital preservation, is an important outcome as it clarifies the *relative* importance that users attach to it.
2. The language used in this report reflects the language of the users. Information specialists, archivists, librarians, etc. have their jargon with specific meanings for words such as information, records, collection, reliability, etc. Many users don't use these words in the specialist meaning, but in a broader, more general way. As this report is intended to reflect user issues and requirements, the users' wording has been preferred.
3. Some of the users were working with digitised (not digitally born) material. Some concerns and requirements reflect the use of digitised material. It is possible that these users are not yet aware of other, more specific issues related to digital born information and documents.

2. Explanation of terms

Assess

Process through which a user determines whether a record, document or data set can be used as it is. Assessment in this sense encompasses a judgement of the trustworthiness (as defined in diplomatics) of a record, document or data set. Depending on the case, this judgement could include an assessment of reliability, authenticity and accuracy (as defined in the InterPARES2 project), or a combination thereof. Users assess digital material both on the basis of the content and formal indicators, thus combining the various concepts.

3. Introduction

A digital preservation plan should consider the needs of its users. A preservation planning system, if designed well, should accommodate these needs wherever possible. These requirements can be determined through qualitative research, and placed into a model for incorporation into a software system.

To attempt a general model of user requirements is challenging. With a vast user scope, including archival, library and data centre users, it will be impossible to determine every potential user requirement, especially given the varieties of subjects and different types of information that could be preserved. Still, through a qualitative methodology, one can determine common user requirements that will be applicable to a large cross-section of users.

The purpose of the second iteration of the PP/3 study was to construct a preliminary version of this model. This model offers a set of selectable user requirements for the Planets Preservation Planning Process. The model follows a tree structure, comprised of general, "root" requirements and more specific sub-requirements as "branches" underneath. This format benefits requirement selection at the planning stage as it allows users to "opt-out" of unnecessary sub-requirements when the parent node in the requirements tree is not selected.

A methodology to approach the qualitative collection of user requirements was developed in the initial iteration of the PP/3 workpackage. This methodology is detailed in depth in the article 'Considering the User Perspective: Research into Usage and Communication of Digital Information' (Snow, et al 2008). The methodology uses the 'data probe' approach to observe research activities of a group of users over a period of time.

After initially testing the methodology in the first iteration, PP/3 selected users of libraries, archives, and data centres for the second iteration. By selecting various types of users, we sought distinctions in requirements between these three user groups. The research was conducted over five weeks in May and June 2008 and analysed in Århus, Denmark in late June. Each partner selected three participants from user groups of libraries, archives and/or data centres.

The affinity analysis procedure, detailed in the D-Lib Magazine article, was conducted in Århus on information from daily input in diaries and statements from user interviews. This analysis produced many potential requirements for the model, in terms of common requirements that applied to all three user groups. The first version of the model was constructed in Århus and later refined through collaborative effort. Some distinctions between users of libraries, archives, and data centres emerged, most commonly in how different requirements would be chosen (or not chosen). Many requirements that were identified and considered as very important by users are outside the scope of the preservation planning process, such as requirements about searching (see section 3.1.1).

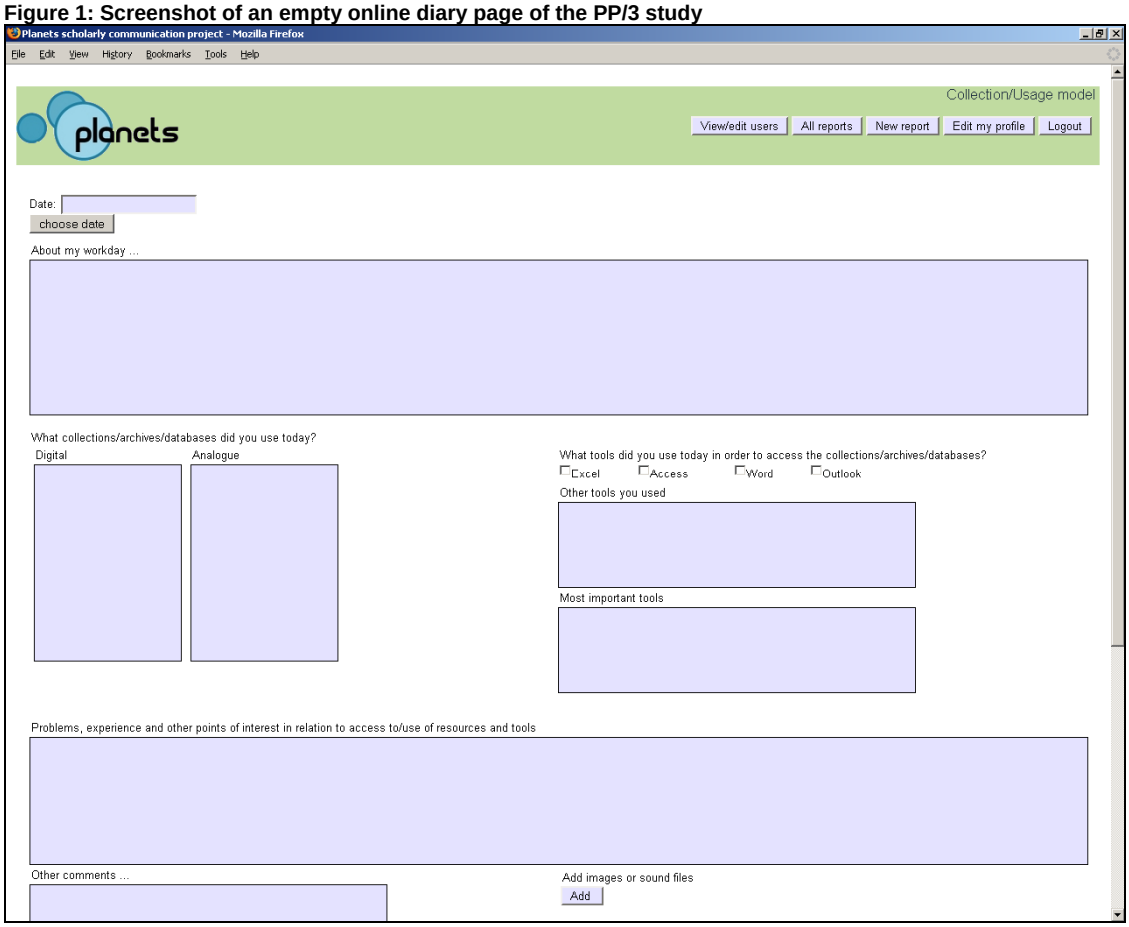
The model is designed using the FreeMind software, an open-source tool for creating mind maps. FreeMind mind-maps can be directly imported into PLATO, the Planets Preservation Planning tool.

4. Methodology

The methodology used in this iteration is both an elaboration and refinement of the methodology used in the first pilot iteration of PP/3. For reasons of clarity, the main characteristics of this approach are described here briefly. For a more extensive explanation, we refer to deliverable 1 of this work package (PP/3-D1) and the article published in D-Lib Magazine of May/June 2008 (Snow, et al 2008).

The methods used in PP/3 are a combination of qualitative approaches including contextual design and data probes. Contextual design is a team-based approach that allows gathering, structuring and analysing of the work habits and concerns of users. The data probe approach is a methodology developed to gather information and is thus used in the first phase of the overall contextual design methodology. We have created a data probe that is specifically tailored for PP/3 purposes. One of the strengths of the data probe is that participants are encouraged to make notes of what happened during a day without interference or guidance of the researchers.
The data probe is a flexible tool that can be designed in various formats. In PP/3's pilot iteration, the data probe consisted of a diary (with daily entries) or an activity map, statements that allowed participants to voice their opinion, the opportunity to hand in screenshots, and the opportunity to hand in files. For reasons of consistency, the three partners aligned their methodology in iteration 2: all used an online diary that was developed in work package DT/7 and slightly changed for the purposes of PP/3 (figure 1).

Figure 1: Screenshot of an empty online diary page of the PP/3 study



As in the pilot iteration, the diary period was initiated by an introductory interview during which the researchers got a good sense of the participants' background, research and interests.
During the four to five week diary period, each participant received four somewhat provocative statements to promote reflection on specific topics.

After the probe period, participants were invited for an interview in which additional information and explanations of the diary input were gathered. The interviews were semi-structured: some questions recurred in all interviews; some questions were personalised and based on the input of the participant during the probe period. On the basis of the answers to the statements and input in the diary and interviews, all three partners created sheets with central critical quotations and syntheses of statements as voiced by the participants. These sentences were used for the collaborative affinity analysis (figure 2).

Figure 2: Partial view on collaborative affinity analysis

During the affinity analysis, conducted in Aarhus on 23-24 June 2008, critical statements were grouped according to common themes. Patterns emerged, statements were regrouped, and from these groupings, requirements were abstracted.

On the basis of the groupings, themes and abstracted requirements from the affinity analysis, a preliminary usage model was designed. This preliminary model was refined, reorganised and elaborated during the weeks after the collaborative analysis. The usage model is the main outcome of the second iteration. It is designed to capture all identified user concerns and preferences for using digital material.

In comparison with the previous iteration, work package participants made the following changes in the methodology:

- Interviews, both introductory and final, were semi-structured, and included some standard questions that were asked to all participants.
- The semi-structured online diary format was used by all participants.
- All participants completed the diary for at least four weeks.
- Statements were more uniform, but with some space for customisation.

In addition to these changes, most recommendations that were made after the first iteration were integrated, including:

- Number of participants per institution was at least three.
- Collaborative affinity analysis work was better prepared and made more consistent so that data were easier to compare.

5. Description of main results of second iteration

While the participants in the study were enthusiastic about the interview topics, the emphasis of the interview was on their own requirements, regardless of whether these requirements were applicable to preservation. As a result, the majority of requirements are relevant for how preserved material is presented to the users, what it should include and how it should allow manipulation.

These requirements are not necessarily valuable for a 'master copy' of preserved material, but can still be elements in preservation planning and in the design of a preservation system. Therefore all results are included in this report. As participants were not primarily concerned with preservation, they frequently used terminology that is at-odds with accurate archival and library language.

After the first pilot iteration, some themes were identified as important to all or some participants (Snow, et al 2008). These themes included:

- Access
- Search and discovery
- Digital versus analogue/copy versus original
- Communication
- Context
- Personalisation
- Backup/data loss

Figure 3: Results of affinity analysis for theme 'Retrieval'

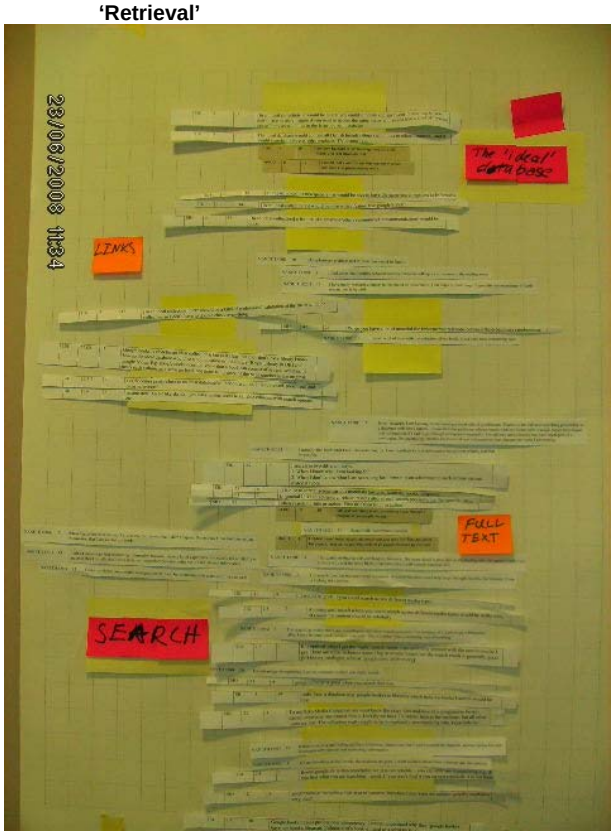
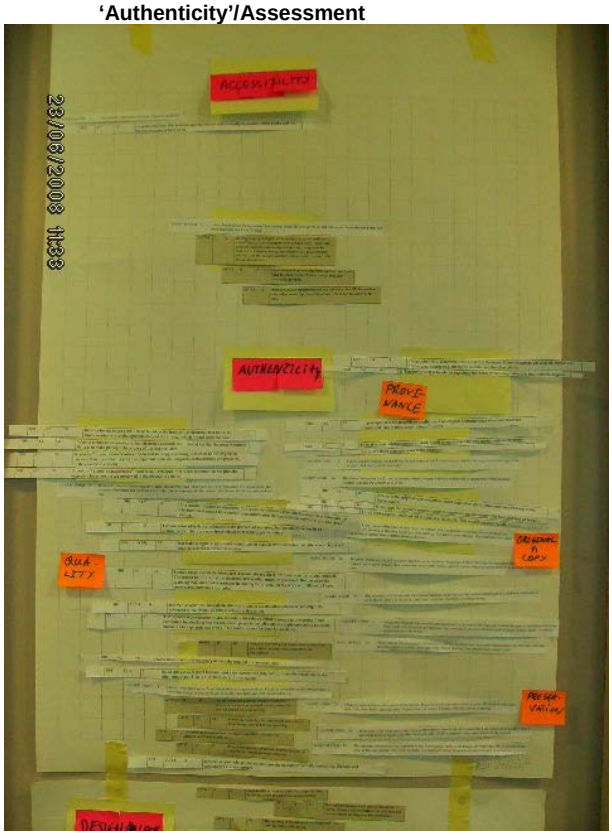


Figure 4: Results of affinity analysis for theme 'Authenticity'/Assessment



Central themes identified during the second iteration were:

- Retrieval (figure 3)
- Depth – layers - granularity
- Flexibility
- Authenticity/Assessment (figure 4)
- Accessibility

- Purpose
- Stage of use
- Large files

Some of the themes recurred; other themes were identified as a subtheme under a new theme, e.g. personalisation under themes 'retrieving' and 'flexibility'; some themes disappeared as their concerns were integrated better into other themes (such as 'Depth – Layers – Granularity').
A major point to note here is that general issues like searching and accessibility are extremely important –if not the most important– concerns of users.

5.1 Analysis of themes

5.1.1 Retrieval

Although retrieving and searching for information does not fit into a model for the preservation planning process, they were identified by the users as one of the most important, if not the most important issue(s) during the interviews and diary period.
Searching is a multi-level and multi-layered issue. It is not only about searching in one collection; it is also about searching and finding relevant information throughout collections and within single documents. In addition, users' expectations about the level of detail of information vary depending on the specific needs of the moment; sometimes they need low-level, detailed information, while sometimes general information about a topic is sufficient. In general, experienced users have personalised strategies to successfully complete their search process, and users seem to be satisfied with their search capabilities. In new situations, the search process is hindered by a lack of knowledge about possibilities and limitations of searching. For some search engines, users realise that the results can only be used "positively", i.e. if they find what they needed. In the case the search engine didn't retrieve relevant results, it doesn't imply that there is no relevant material.
Users have high levels of expectation and bad search capabilities (in software) can partly influence research activities. This issue was especially identified by archival and library users who voiced concerns that some collections are more accessible – meaning better described and thus easier to search in – than others. This results in a bias/preference for collections that are more accessible. As such, the quality and depth of archival descriptions and thus search results shape research results; researchers obviously prefer sources that are easy to consult because it allows for quick results in a virtual world collection with ever increasing numbers of sources/collections that are possibly of interest for their research. One example was provided by a user who illustrated how he retrieved information he would never have found without digitally accessible collections. He accessed a database with digitised newspapers because colleagues had informed him that it allowed for full text searching, thus allowing for quick results. It resulted in a number of interesting newspaper articles that added to the work of the researcher. The whole process had not taken too much time because the searching facilities were easy to use.

In addition, users indicate that some form of serendipity should be possible during the search process in order to allow for interesting search results that would otherwise not be presented. This may be something like the Google 'I'm feeling Lucky' button, or merely search results that appear above or below the relevant search results – results that may open up new directions despite not being exactly what a user is looking for initially. In the paper world, the physical placement of books and records allowed for searching through related items. In a library, users could also take a look at the books that were on the same shelf with the same theme. In an archives, records are grouped in archival boxes, and in many cases, users get all files in a box (although they only requested one) and are thus able to check the other files in the box too. Users indicate that they sometimes find additional interesting information by doing this, and thus would like to have this capability in an electronic environment as well. Search engines that indicate related material are favoured by some users.

Full text
Many users mentioned the importance of being able to search the entire text of a document as opposed to just an abstract or metadata. In libraries and archives, this was most common as the scientific data found in the data centre users was not primarily text-based. Though most electronic

formats allow for full-text searching, it is still an important requirement that will be necessary in any user-defined preservation system. For large text files that are not described in detail, the ability to search full-text is even more important.

5.1.2 Accessibility

In this study, respondents overwhelmingly emphasised the importance of accessibility. This theme actually filtered into all of the other themes, such as *retrieval* (as discussed above), or *depth – layers – granularity*. The phrase “not time consuming” was repeated by several users. It is clear that users want digital system to be beneficial to their work, rather than an impediment – a common sense view, of course – but one that is still explicitly expressed due to existing flaws in information systems.

Accessing content
The act of accessing information or data is at the core of all electronic resources, and that access is often as simple as a mouse click. Though this process is often transparent, it is central to any system. Once electronic records are retrieved, it is important that there are no impediments to users actually accessing the content of these records.
In the construction of the preliminary model, where requirements are organised into a tree structure to allow selection in a preservation system, access envelopes the majority of the requirements.

Technical problems
Users cited technological problems as one limitation to their work. Researchers using scientific data indicated that enormous data sets could be cumbersome to work with due to limitations on memory, storage and display areas. An essential stage of their work involves reducing raw data to its essentials for their own needs; these reductions of course vary by discipline. One user in particular stated that his process of reducing raw data was personalised with custom-written software, yet it was still very time-consuming. The large set of raw data must be preserved for matters of scientific accuracy and to allow peers to reproduce results; thus, the storage required for his discipline (astrophysics) is much greater than if it was merely storing “human readable” information.
Similar observations are valid for archival users. Historians use records as material for their reconstruction of history, but these records can be, for instance, very long (text) documents. These documents are not problematic because of their size in bytes, but they pose similar issues: it is difficult to store them as they are, and it is time consuming to extract the information they need for their research and to organise this information in a way that is meaningful and easy to retrieve.

Reading texts and documents
Users also indicated that while they prefer digital materials for their work, they still frequently print copies. Dependent on the situation, they prefer a digital or a paper copy. The paper copies are often used because they are easier to read.

5.1.3 Depth – layers – granularity

Users combine several actions while working with information: they search, access, analyse and absorb information simultaneously. During this ongoing process they constantly make assessments about the usefulness of information and its specific place in their research. Information can take various positions and can be used for many purposes. It can be final outcomes of the research process, but it can also be information that is necessary for further investigations, or intermediate research results. Depending on the particular situation, users will sometimes need an overview of what is available, sometimes they’ll need detailed information on the basis of which they can make decisions that further guide their research, sometimes they’ll want to get a general idea of how a specific type of information looks like, etc. For users, all of these various examples require information to be searchable and accessible on different abstraction levels. These levels of abstraction may be compact or concise (for instance as an abstract), or very detailed. Users prefer that information is available at the abstraction level that they need it, without having to handle or re-structure it for optimal use in their context.

5.1.4 Flexibility

Users require digital materials to be flexible. The ability to re-use and recombine information is necessary to the construction of one's original work. Users expressed these needs by emphasising the importance of their unique, personal approach.
The initial question when accessing electronic information is that of how it will be used. If it is something that only needs to be referred to – a "read only" bit of information – then users are much less demanding in what they require. If this is the case, then flexibility requirements are unlikely to be any more complex than being able to cut and paste quotations.
If a user intends to incorporate material into their own work, then it raises an entirely new scope of requirements. Users may then need the ability to edit or annotate the information, convert it to different file formats, or conduct any other forms of transformations. Again, the underlying requirements for accessibility (see section 5.1.2 above) apply to flexibility – users want things to be easy, and not time-consuming.
Scientific data often comes in formats which are not always as flexible as formats that are of more general use. Users may experience difficulties using these systems and often spend more time than they wish trying to convert files and retain the essential properties of the data. Standardised formats for scientific papers, such as TeX (for typesetting) and PDF (for publication), greatly improves the user experience – yet there are still sometimes lengthy processes involved in migrating data before writing these reports.
An additional issue is that information created in the past is unlikely to be structured in the way that is fit for immediate integration in new research. Therefore, users have to invest a lot of time in transforming the structure of the information so that it is ready for use, analysis and interpretation in their research.

5.1.5 Authenticity – Trustworthiness – Assessing information

The value of information is not always clear. Users of libraries, archives and data centres all use different means to address this, and their interpretation of what authenticity-trustworthiness entails is rather fluid. For some users it is about reliable sources, while for others it is about the accuracy of the information. Often, a user shifts the importance of these notions from one to another, in order to clarify whether information is 'trustworthy' or not.
However, users generally trust information that is preserved by libraries, archives and data centres, and unless there are clear reasons to doubt, the trustworthy character of information preserved by libraries, archives and data centres is accepted at face value.
More as a natural reflex than as an attitude that is shaped through education, users will assess information on the basis of sets of requirements that are (in their opinion) appropriate. This assessment makes the assumption of authenticity and trustworthiness plausible.

Some users indicated that more formal procedures of assessment and approval of electronic material would be beneficial. Published material has already passed one stage of 'assessment'. The system of peer review that guides academic work is evident in the content of the information. One user indicated that also for collections there should be a kind of professional validation of the material so that users need not double check everything. In the peer review system, readers can trust the assumptions and hypotheses that are made, and also the information on which the article is based upon, can be accepted, as the professional validation has been made.
Users of archives, in dealing with unpublished records, base their assessment of the information on the reliability and accuracy of the information as described in the records, and if necessary on the knowledge they have about the author and broader context of the records. They look for things like consistency and logic in the content to determine whether they accept the information or not. If the content is suspicious, archival users look for other elements that can validate or invalidate the record's reliability. In this case, the assessment becomes a stepwise approach in which users try to find decisive evidence or plausible assumptions that clarify the status of the information contained in the records. Users of data centres may use either of these means. Some data sets originate from highly recognised research programs or institutions and are widely accepted as authoritative and are thus considered as trustworthy.

Users expressed a strong need to preserve this multifaceted trustworthiness of preserved material. The reputation of a journal or publisher is often good enough for many users; therefore some indication of this in a preservation system is necessary. Metadata that carries the name and issue

number of a journal would be suitable for many users. Additionally, the physical look of a document may convey that it is part of that journal, though most users are not concerned with the design and layout if there is some other indicator of trustworthiness.
However, users generally want the look and feel to be preserved in the case of *digitised* information. Users are afraid that in the process of digitalisation, information is lost. Retaining the original look and feel (by offering a photographic reproduction) ensures users that all information has been preserved.

5.1.6 Purpose – Stage of use

During the research process, users simultaneously execute various actions, with different purposes. Some of these research actions require general information; others are best served with detailed information. Depending on the purpose – or the stage of use – users have specific requirements for information. These themes of purpose were incorporated into the Flexibility section of the model.

5.1.7 Large files

Large files present users with two main problems.

1. The size and scope of the files make it harder to manage them compared to smaller sized files. They take longer to download, and scrolling through them may be slower than in shorter documents. It is more difficult to attribute large files to one specific folder in the classification system of the user because the content of a large file may covers multiple themes that span the user's own organisation system.
2. Also, the content of large documents (files) is usually not described with as much detail as documents of a shorter length. As a result, the content of large files may be more obscure. As an example, descriptions in an archival institution will indicate the overall content, but not the content of specific chapters or other relevant themes. Because large files are generally not as well described as shorter files, they are less 'attractive' to users. It is more time consuming to assess whether they contain useful information, and it is also more difficult to refer to this information in the large file/long document. In general, users have to invest much more time in these large documents than if the same information was available in multiple, smaller documents.

6. User model

The development of the model of user requirements progressed through several stages. In the following sections, we describe the evolution of the model as it progressed from the affinity analysis stage to the completed stage This model (figure 7), while the 'final' model for this iteration of the project, is still a preliminary model and will be developed further in the future iteration.

6.1 Development of the initial model

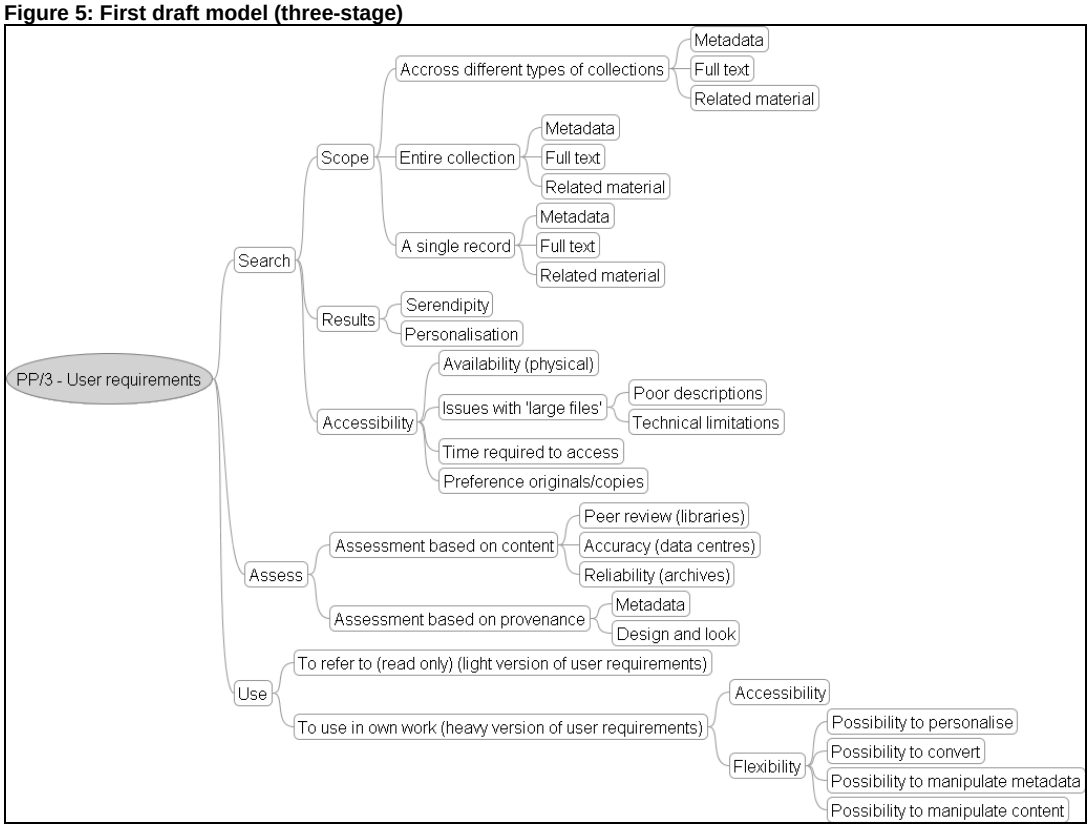
The first version of the model was developed after the affinity analysis. The themes as described above (section 5.1) reflect the original grouping as an immediate result of the affinity analysis. During the model-building process, several drafts were created, and requirements were shuffled and regrouped several times.

This early draft (figure 5) was based on three activities that all users go through:

- search
- assess
- use

The search stage is the stage during which users retrieve information that seems or is relevant for their work.
The assessment stage is a more implicit stage in which users check whether the information they are accessing is truthful. In most cases, users trust what they read unless there are 'obvious' reasons to doubt.

Figure 5: First draft model (three-stage)



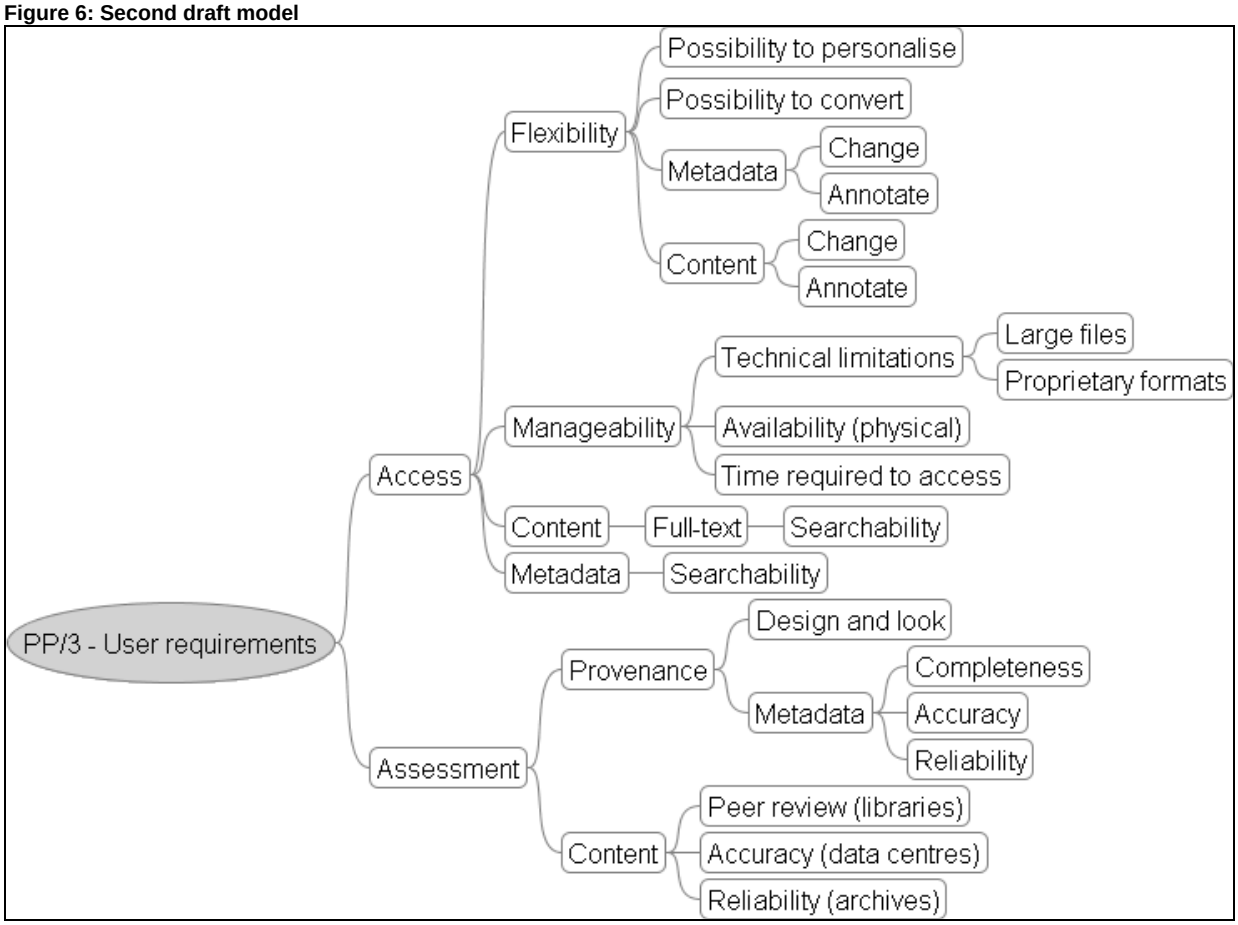
The use stage is the stage in which users want to work with the information they have found. Depending on how they will use the sources they have found, users have different expectations on

what they want to do with the copies that they are accessing. These themes are described above in section 5.1.4.
This first model has the advantage that it shows the sequential (although partly overlapping) stages that a user follows, and that it shows the (general) requirements that are relevant at each stage. The main disadvantage is that it doesn't necessarily group requirements in a logical structure that is useful for the more general purpose of preservation planning.

6.2 Refinement of the model

The requirements, presented for use in the Planets Preservation Planning Tool (PLATO), are presented as a series of choices. The tree format is perfect for the decision-making component of the system; if one requirement is not needed, then its subrequirements will not be considered.
Many of the requirements expressed by the users were rendered irrelevant by the nature of preservation planning decision-making. For example, the "use" branch of the first version of the model, while a perfectly valid way of conceptualising the use of digital materials from a theoretical point of view, is out of the scope of a preservation planning tool. If a user requires the ability to annotate a document's metadata, then that can be directly translated to a presentation rule.
Therefore, an improved draft was developed that integrated the requirements that are present in figure 5 in a more structured manner for preservation planning purposes. It is built around two main branches: *assess* and *access* [or *use*] (figure 6).

Figure 6: Second draft model

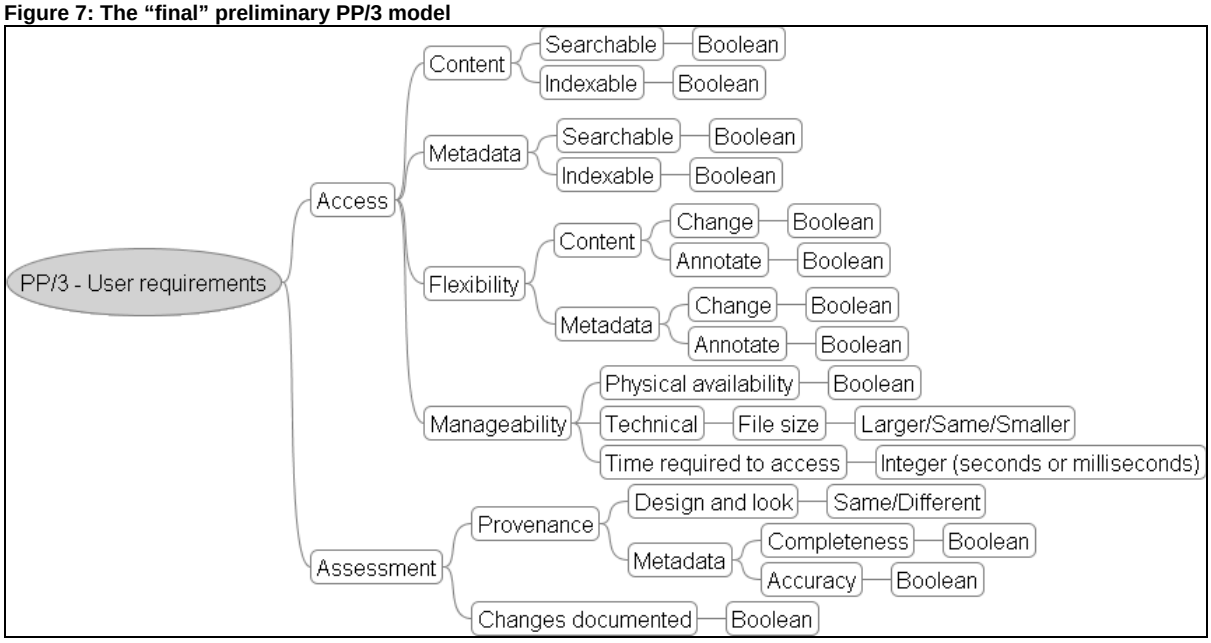


6.3 Completion of the preliminary model

After further collaboration among the PP/3 team, the preliminary model was completed. This model was similar to the second draft, but clarified a few more requirements in terms of relevancy to the preservation planning process while also returning some requirements that had previously been discarded.

Additionally, units and measurements were added to each node of the tree. The majority of these were Boolean (true/false) but in some cases, a quantitative measurement was needed.

Figure 7: The "final" preliminary PP/3 model



6.4 Description of the model

Access
All users require access to information. This access is what gives purpose to preservation systems. Regardless of whether someone is using a library, archive or data centre, there will be an interaction with information. These requirements are organised into four nodes to cover the areas of content, metadata, flexibility and manageability.
Below are brief explanations of each node, organised hierarchically. If some requirements seem repetitive, obvious, or unnecessarily explicit, please remember that they are attempting to describe all possible outcomes.

Access > Content

The content branch describes requirements for accessing the content of digital information. This is distinguished from metadata.

Access > Content > Searchable

For textual information, a user may require that the entire text of the document is searchable. The ability to search through the full text of a document may be present in some file formats and not in others; therefore it will affect the presentation of information from a preservation system.
For non-text content, such as an image or raw scientific data, this requirement may not be applicable as keyword searching would be covered in the metadata requirements below.

Access > Content > Indexable

Though originally it seemed that the multitude of user requirements regarding searching would exist outside of the scope of a preservation planning model, the ability for content to be indexed by a search engine or database is essential to the needs of users. Thus, a yes/no flag on whether the content can be indexed (which may vary depending on format) is required.

Access > Metadata

The metadata branch describes requirements relating to accessing metadata. Some users may not require metadata at all, although this is unlikely. Which metadata should be preserved for users will of course depend on the individual user; no general assumptions can be made. In the assessment phase, it may be possible that users assess the trustworthiness on the basis of provenance (as registered in the metadata). For access, the presence and searchability of the metadata is most important.

Access > Metadata > Searchable

Similar to the Access > Content > Searchable requirement, a user may wish for the metadata to be searchable.

Access > Metadata > Indexable

Similar to the Access > Content > Indexable requirement, a user may wish for metadata to be indexed by search engines and/or databases.

Access > Flexibility

The flexibility node describes requirements for how the user works with a document. If a user merely requires a document for “read-only” purposes – for example, just to include as a supplement to their work without editing it – then these requirements will not be selected. Users indicated that they want to include information from documents, records and data sets in their personal information system; therefore, it’s handy if the format and structure of these retrieved documents, records and data sets facilitates easy handling, changing, copying, regrouping of data, etc.

Users generally spend a lot of time processing information and data, so the ease of handling of retrieved documents, records and data sets could result in substantial gain of time for the researcher.

Flexibility includes (as specified below):

- Documents, records or data sets that are easy to copy in a work document of the user.
- Data sets and documents are easy to use because they are in a format that enables easy migration or conversion to the format that is used by the user
- Data in data sets are easily restructured so that the grouping and structure of the re-used data corresponds to the needs of the user.

Access > Flexibility > Content

These requirements cover the user's ability to alter content, following the themes of flexibility. The abandoned “stage of use” tree (found in the earlier version of the model [figure 5]) is expressed here through the subrequirements that deal with editing and alteration.

Access > Flexibility > Content > Change

A user may require the ability to change the content of a copy of a document, in which case a format that allows editing will be required.

Access > Flexibility > Content > Annotate

A user may require the ability to annotate content, in the form of comments or notes, without altering the original content.

Access > Flexibility > Metadata

These requirements cover the user's ability to alter metadata, similar to the main content flexibility described above.

Access > Flexibility > Metadata > Change

A user may require the ability to alter metadata.

Access > Flexibility > Metadata > Annotate

A user may require the ability to annotate metadata without altering it.

Access > Manageability

These requirements describe the user's interaction with the information from an organisational point of view. They are in response to concerns that digital information is often difficult or unwieldy to work with.

Access > Manageability > Physical availability

Many users express the need to create a physical copy of their data to aid in their work. Users (especially those using texts) indicate that they prefer a material copy to read as reading from a screen is considered tiring and more difficult, especially if longer (parts of) texts are to be read. The digital documents should allow good quality and easy to read documents.

Access > Manageability > Technical

These requirements describe requirements relating to hardware and other technical matters. Though file size is currently the only requirement under this tree, we foresee further technical user requirements in the final version of this model (after iteration 3).

Access > Manageability > Technical > File size

Some users have indicated difficulty in working with large data sets, images, etc. Computers have limitations in memory and storage, and massive files may fill the space. Additionally, high resolution images may be difficult to work with due to the size of the video display. This requirement is measured as a field of comparison – the resulting file size may be larger, smaller, or the same size as the original file.

Access > Manageability > Time required to access

Throughout all areas of the study, users emphasised the value of their own time. The phrase "not time consuming" was repeated multiple times when discussing almost everything. Clearly, no one is going to be satisfied with systems that cannot retrieve and open data sets, documents or records in timely fashion. Still, it should be a guiding principle of systems design to simplify procedures and provide the user with an easy to use interface. The time required to access a preserved object can be measured and used to rank possible outcomes of a conversion or migration. This can be specified in integer format, as number of seconds or milliseconds.

Assessment

Users require some indicator of trustworthiness. This issue of trustworthiness varies wildly and may be different for every user. PP/3 has determined that – from a user's point of view – this assessment can take place on the basis of two means, an assessment based upon content, and an assessment based upon provenance (as part of the context in which a data set, document or record is created). Provenance may be recorded in metadata, or it may be communicated through non-content means such as the design of a document and the presence of formal characteristics in a document. A publication may use a standardised font and layout, and this may indicate the source to a user.
The issue of content-based assessment is much more difficult. The PP/3 study examined users of libraries, archives and data centres and discovered differences in how each assigns 'authority' to information. Information retrieved from libraries is already published; therefore it has already undergone peer review so authority is not a problem. Users of archives, working with unpublished material, have reliability and accuracy as their main concerns.(See also Hedstrom et al., 2006) The data centre users had several approaches to assessing material, though often the data centre itself was considered to be an authority.

Assessment > Provenance

These requirements are for a provenance-based method of assessing information.

Assessment > Provenance > Design and Look

These requirements are for provenance as identifiable through the design and look of a document. Though few users will require exact reproduction of fonts, colour, and layout, it may be required for some purposes such as historical reproduction.

Assessment > Provenance > Metadata

This requirement specifies the inclusion of metadata that records the provenance of information.

Assessment > Changes Documented

Initially, we discovered that many users used the content of the information as a means of assessing it. For example, an archival user may check the correctness of the content of archival material to determine if it is of value. This did not translate into the framework of PLATO, though an alternative method for content-based assessment was discussed.
As documents are altered, a revision history can be recorded in digital formats. This is perhaps most commonly seen in computer language versioning systems such as Subversion or CVS, though a similar approach can be taken by a Microsoft Word document using the "Track Changes' feature. This can be measured in varying levels of detail, which will be investigated further in iteration 3. For now we have included this as a Boolean field, to specify if changes are documented or not.

6.5 Other requirements

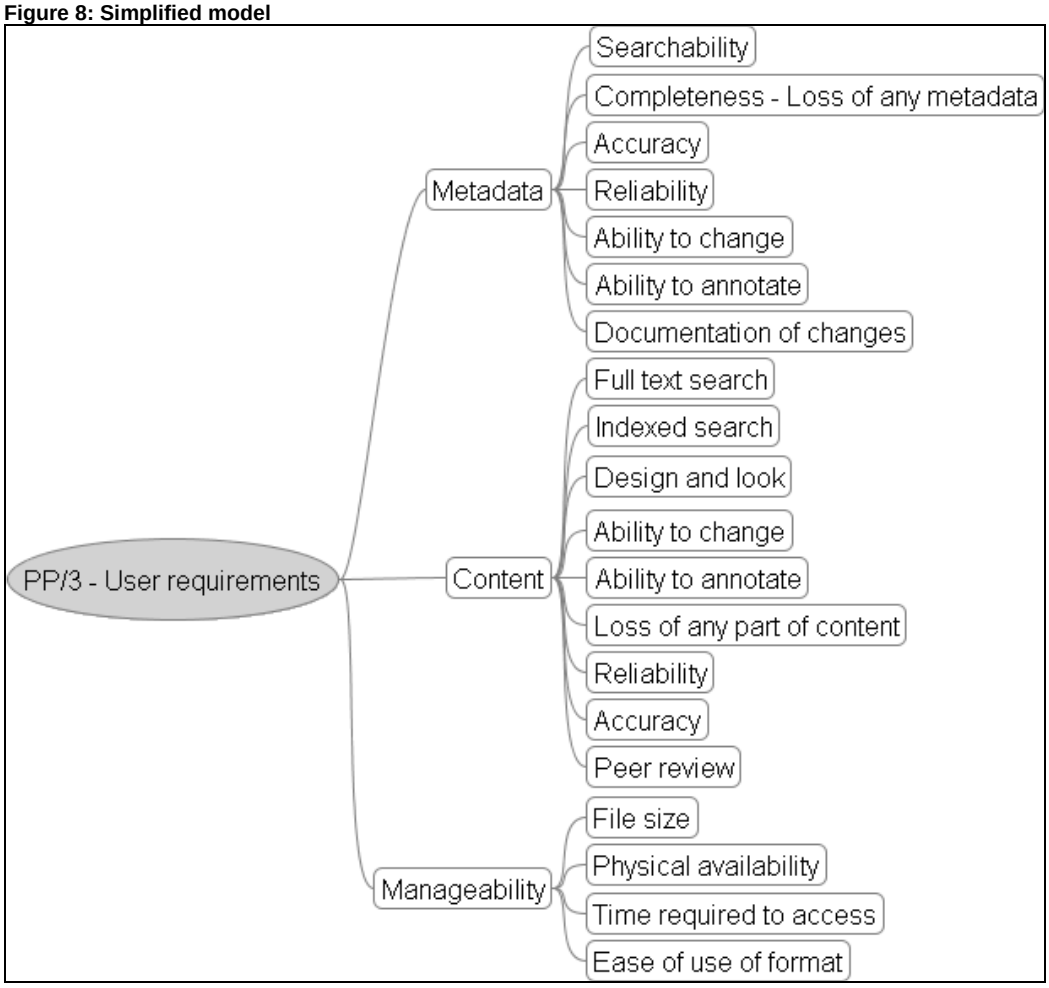
The requirements removed from the PLATO model are still of general interest despite lying outside of a function in PLATO. Because these are "user requirements" and were uncovered through our studies, they are described here. These descriptions are intended to be guidance; we hope that they will influence the design of preservation systems from a user point-of-view.

The topic of searching is continually returned to throughout the user studies. The search is the root of all information retrieval, whether the information is stored in libraries, archives or data centres. Users continually expressed their views on search systems, and stated many requirements for finding information.

Three different scopes for searching were revealed. The first scope is the single record – users expressed the requirement to search within a single record. This requirement appears in the model as *Access* > *Content* > *Full-text* > *Searchability*, which can be subdivided into different types of search (keyword, phrase, complex, etc.)
The second scope is that of multiple documents – users expressed the requirement to search across a set of documents, and the third scope is that of multiple collections – users often want to search for information across multiple collections, and not necessarily collections of the same type of information. These scopes are not related to the preservation objects, but more to the system in which they are grouped / preserved. The ability for a digital object to be indexed and therefore found in a search function has been incorporated into the model as *Access* > *Content* > *Indexable* but there is no way to indicate how an object can be found across several collections.

The idea of 'serendipity' was expressed several times in the study. Inherently immeasurable, serendipity leads to new breakthroughs in work and is a by-product of the search process. In terms of preservation, this would involve the relationship of a document to other documents, discovered by the same search process that the original document was found in.

Figure 8: Simplified model



7. Conclusions

7.1 5.1 Preliminary conclusions

The user study as described above has revealed that – if not asked explicitly – users are only mildly concerned about requirements that could be posed upon digital preservation. In their overall research activities, users' *main* concerns are searching (and finding) relevant information, accessing that information, and ease of use of that information.
However, some user requirements are about specific characteristics of the documents, records or data sets. For digitised, non-native electronic documents (which some users were working with), these concerns are extremely relevant as the quality of for example scanned archival (hand-written) documents is often not optimal. Also, in the process of scanning, some information may be lost (for instance notes in the margin), hence the users' concern about loss of information.
In an environment with born-digital documents, it is expected that users still focus on the importance of excluding or minimising loss of information, but due to its nature requirements may shift to other characteristics. For the hybrid paper-digital copy world, users indicate that the original paper original was sometimes double-checked against the content of the digital copy. Procedures in which the process of checking the quality of copies would be executed by an authoritative source, could replace this individual time consuming process of double checking. Whether such formalised processes are needed in an electronic environment, or are replaced by automated procedures, is yet to be researched.
The fact that most of the users in this study were working with originally paper documents may distort this result, although other studies have shown that loss of information is a general concern of users (Casey and Jansen 2003).

The purpose of retrieving information by users, whether in an archives, libraries or data centres, is to use it. Users formulated many (general) requirements which indicated that ease of use (easy transformations, conversions, copy and paste operations, etc.) is a key requirement in their daily work. As these activities are very time consuming, users try to gain time by using more advanced tools or focus on collections that are easily accessible. These user requirements do not necessarily reflect specific needs vis-à-vis preservation of the material, but are often closely related.
On the basis of the three stage model, it is clear that the opposing requirements about preserving the material as close as possible to the original (for assessment; with look and feel as close as possible to the "original"), and ease of use/flexibility (for use; in a structure and format that allows easy editing, copying and handling) are challenging memory institutions.
In addition, although users do not necessarily mention metadata, it is required to assess the provenance or trustworthiness of the data sets or documents. The availability of metadata that allow the various assessments of users (reliability of information, correctness of what is described (records) or measured (data sets), etc.) and retrieval of the information, is one of the main derived themes.

7.2 The next iteration

This iteration of the research has resulted in a set of general user requirements that transcend issues that are relevant for digital preservation. In the final phase of PP/3, it will be essential to examine the general requirements in more detail. We will retain the qualitative approach, but look to directly observe users interacting with preserved objects. This way, these requirements can be directly evaluated in an actual usage situation rather than a theoretical one. It is our desire to refine and improve this model, producing a final version at the end of the PP/3 workpackage.

References

- A. Casey and B. Jansen, *National Archives of Australia AtoR Preservation Project. Requirements Analysis Report*, s.l., 2003.
- M.L. Hedstrom, C.A. Lee, J.S. Olson, C.A. Lampe, "The old version flickers more": digital preservation from the user's perspective, *The American Archivist*, 69, 1, 2006, pp. 159-187.
- Ontology C of InterPARES 2 project. Available at: http://www.interpares.org/ip2/display_file.cfm?doc=ip2_ontology.pdf
- John Roeder, Philip Eppard, William Underwood and Tracey P. Lauriault, "Part Three – Authenticity, Reliability and Accuracy of Digital Records in the Artistic, Scientific and Governmental Sectors: Domain 2 Task Force Report," [electronic version] in International Research on Permanent Authentic Records in Electronic Systems (InterPARES) 2: Experiential, Interactive and Dynamic Records, Luciana Duranti and Randy Preston, eds. (Rome, Italy: Associazione Nazionale Archivistica Italiana, 2008). Available at: http://www.interpares.org/display_file.cfm?doc=ip2_book_part_3_domain2_task_force.pdf
- K. Snow, B. Ballaux, B. Christensen-Dalsgaard, H. Hofman, J. Hofman Hansen, P. Innocenti, M. Poltorak Nielsen, S. Ross, J. Thøgersen, Considering the user perspective. Research into usage and communication of digital information, *D-Lib Magazine*, vol. 14, nr. 5/6 (May-June 2008). Available at: http://www.dlib.org/dlib/may08/ross/05ross.html (accessed on 26 August 2008).