The Use of Metadata and Preservation Methods for Continuous Access to Digital Data

Groenewald, Ria

Dept. of Library Services, University of Pretoria

Breytenbach, Amelia

Jotello F. Soga Library, Faculty of Veterinary Science, University of Pretoria

Abstract

Purpose - The purpose of this paper to investigated the awareness about digital preservation and what must be done towards preserving valuable original digital material. The paper also discusses the use of metadata principles and the implementation of tools for the preservation of documents stored on personal computers.

Design/methodology/approach - Data loss prevention starts with the creation of a digital object. However, methods to minimize the loss of digital data are often ignored, the use of metadata structures embedded in digital objects from the outset thereof are recommended as a starting point towards good preservation principles. The need to create awareness on the issue of digital preservation was promoted by the authors at various occasions during 2008, as the number of incidents of data loss and costs involved continue to be of concern to all involved. Whether the loss occurs by a malicious attempt, or an inadvertent mistake, it can be diminishing either personally or to the institute/company where it occurs. Data was collected through a digital questionnaire and lliterature studies were done on several strategies, policies and best practices. Personal visits to libraries actively working on digital preservation and curation formed a basis for conducting the study. The implementation of tools that can be used to preserve and retrieve digital content by individuals was also investigated.

Findings - Digital objects should be archived with metadata about the object and the creation thereof. Metadata need not necessarily be structured and controlled when used by individuals or small groups for preservation of self owned data. The metadata content however, should describe the object, the method of creation and technologies used in the creation. All changes to the document should be captured in the preservation metadata. Future access to digital content does not only depends on one preservation method but on a sequence of strategies and methods applied to the digital content.

Originality/value - The paper provides a broad overview of certain aspects that must be taken in consideration when implementing digital preservation strategies. The value of metadata need to be widely accepted and implemented and the use thereof promoted to creators of personal digital objects.

Type of paper - Case study

Keywords - Preservation metadata; digital preservation; data curation; digital data; preservation strategies; computer data

Introduction

Libraries, museums and archives traditionally are the custodians of valuable artefacts and information. These valuables were acquired from individuals and other institutions and stored with well-managed conservation practices to ensure the long-term access thereof. Institutions are now facing new access and preservation issues, as personal documentation are being donated to them in digital format without the personnel at the deposit institution having the necessary skills to handle and store such information.

Descriptive metadata as an application method for digital document access has established itself over the last few years. Preservation metadata methods however, have not yet receive the same intensity of application to electronic documents, although these metadata sets are crucial for the preservation of the document format and to retain the significant properties (look and feel) of a document.

Data loss prevention starts with the creation of a digital object. However, methods to minimize the loss of digital data are often ignored, thus the use of metadata structures embedded in digital objects from the outset thereof are recommended as a starting point towards good preservation principles. The need to create awareness on the issue of digital preservation was promoted by the authors at various occasions during 2008, as the number of incidents of data loss (http://datalossdb.org/index/ largest) and costs involved continue to be of concern to all involved. Whether the loss occurs by a malicious attempt, or an inadvertent mistake, it can be diminishing either personally or to the institute/company where it occurs.

Information created in digital format and selected for archiving needs to be preserved in the format of creation without any restrictions embedded in the document. Documents intended for web use are usually disseminated for fast, easy and clear access through the internet but do not serve the archival purpose. Although dedicated "web formats" also need to be able to migrate and have a sustainable "object life" it does not form part of the focus of this paper.

Social networks have an influence on the flow of digital data. Personal documentation such as photos, life experiences (travel, events and personal opinions) are posted to blogs, wikis and other social web sites such as Flickr (www.flickr.com) and Facebook (www.facebook. com). Libraries now acquire specific places in those web sites for display and presentation of valuable photo content and encourage their viewer via these web sites to add information to the displayed content, i.e. Library of Congress (www.flickr.com/photos/library_of_ congress/). Web 2.0 tools, however, are most often used by individuals as replacements for former diaries, photo albums, letter boxes and filing systems of personal documentation. Metadata in the form of social tagging is usually very basic and descriptive in nature in

these applications, thus neglecting preservation metadata. Web 2.0 tools orchestrate a clear shift to the individual as a publisher, and move the balance of power in relation to information away from the organisation.

During 2008, the authors ask students and attendees of workshops to indicate whether they still have the first photo taken by them on a cell phone, the response was overwhelming negative. This is in contradiction to the paper print of many older "first photos" which are still in existence.

The selection of paper material is normally defined as an artifact that has not been copied or derived (Jantz et al, 2005); this principle should also be applied to digital material. The PoWR Handbook, funded by JISC for web preservation, concentrates on strategies for the preservation of web material. (University of London Computer Centre, 2008). In this handbook the MoSCoW approach to the selection of digital archival material is encouraged for digital preservation strategies. Although the concept is broadly stated, it does provide a valuable starting point for thoughts on digital preservation

M: Things you/institution must preserve

S : Things you should preserve, if at all possible

C: Things you could preserve, if it does not affect anything else

W: Things you won't preserve

These broad principles can be applied in the selection of all digital material for long-term storage and archival purposes by individuals and organizations.

Methods of Study

1. Questionnaire

During February 2009, a digital questionnaire containing questions about file formats, level of knowledge about digital preservation, back-up methods, and type of digital material stored on computers, was circulated to two separate South African Library list-serves. The participants in the research therefore mostly originated from South Africa. The purpose of the questionnaire was to determine the magnitude of awareness about digital preservation and what was done towards preserving valuable original digital material.

The outcome of the questionnaire indicates a lack of knowledge on preservation strategies and digital content management, ignorance on the need to preserve digital content for future access, as well as a need for training in basic digital preservation methods. It further indicates the usage of personal computers (office or at home) and the format types of electronic documentation, stored on private computers. The outcome show a variety of different well known format types, with no indication towards the usage of any unique file formats, i.e. those used for datasets etc. The niche market

of the questionnaire was not researchers but people who work in the knowledge and information sector.

2. Literature studies and visits

Literature studies were done on several strategies, policies and best practices. The information was mostly accessed online. The amount of articles recently published on the topic indicates

- a concern towards digital fragility
- a concern for future access and preservation of digital assets
- the need to create awareness amongst creators of digital content; and a
- growth in study- or working groups on the topic of digital preservation.

Most of the studies cover digital preservation strategies, policies, sustainability and tools that can be used for digital preservation by larger institutions but omit the layman as a creator of valuable digital content. Many smaller institutions do not have the necessary IT support to help with the installation of preservation tools. The Library of Congress' Digital Preservation website [5], however, does contain valuable information for personal digital archiving and provides instructions on the preparation, protection and preservation of valuable digital information. (Library of Congress, 2009)

2.1 The DCC Curation Lifecycle Model

The DCCs [6] (Data Curation Centre) Preservation Lifecycle has been studied. The model comprehends all the stages that have to be done for the successful curation and preservation of data. The stages described are:

- (1) Creation of the object
- (2) Ingest to archival, museum or library server
- (3) Appraise and select
- (4) Normalisation of object
- (5) Assigning of metadata description
- (6) Preservation planning
- (7) Storage for access and reuse transform
- (8) Curation lifecycle management this include disposal, migration and reappraisal

A digital object should to be actively managed at each stage of its life, and preservation strategies need to be implemented from the creation thereof. Originators or collectors of digital material will know what the short-term as well as long-term value of the digital material is and the anticipated life-span there of. This concept must be the starting point for preservation of digital objects. The DCC model describes sequential activities to ensure that all necessary stages are preserved.

2.2 PREMIS Report

Although the primary uses of PREMIS (a data dictionary for preservation metadata) are for repository design, this report has been used as a starting point for the use of preservation metadata. (Caplan, 2009). Preservation metadata can be used by every creator of digital data and should be encouraged (Figure 1).

2.3 Study visits to other Libraries

Personal visits to libraries in the UK and Egypt also create an opportunity to learn more about preservation methods of digital content.

- (1) A visit to the Wellcome Library, UK (http://library.wellcome.ac.uk/) gave important insight in the starting process of initiating preservation methods. Information about assigning metadata to personal documents (demonstrated later in this article) and also the importance of digital wills were knowledge acquired for the study. Digital wills are important because of probable copyright problems that might arise during the preservation life span of an object. It is best to create a digital will beforehand which should accompany the original material. Guidelines and future access rights should be described in the will and it be used for management and curation purposes.
- (2) The Alexandria library (www.bibalex.org/English/index.aspx) hosted a duplicate copy of the Internet Archiver's web captured information. More than one copy of the Internet Archiver's information is stored around the globe for safe custody. The library of Alexandria is also the custodian of other digital material that is archived and managed by them for future access. The practical principals of digital preservation could be observed during the visit.
- (3) The Digital Preservation Training Programme (www.dptp.org), which is operated and organised by the University of London Computer Centre (ULCC) and other leading experts in the field, gave and insight in the OAIS model and provide a chance to network with colleagues on digital preservation and the challenges that digital preservation needs present.

3. Software

During the study the authors investigate the implementation of tools that can be used to preserve and retrieve digital content by individuals, with the emphasis on preservation. Although our focus has been on preservation we realised that preservation practices cannot be fully implemented without a good content management system and search facility. Different types of software have been identified for preservation usage.

- (1) Content Management
 - Joomla (www.joomla.org/)
- (2) Format Preservation
 - Xena (http://xena.sourceforge.net/)
 - JHOVE (http://sourceforge.net/projects/jhove/)
- (3) Web 2.0
 - Web Curator Tool (http://webcurator.sourceforge.net/)
- (4) Search Facilities
 - Windows Explorer
 - Copernic (www.copernic.com/)

As IT support and back-up is most often a problem, the aim was to download tools from open source software without any programming knowledge or external help. Unfortunately, this seems not possible as a certain degree of programming knowledge is necessary for installing and implementing the software. Some of the tools, such as the Copernic search facility, was easy to install and only comprises the downloading of the tool with the installation done automatically by the software self.

The web curator tool is developed for the capture and preservation of web products and need intervention by expertise to install. Creators of web pages should be encouraged to preserve the HTML files on their computer with enough technical and descriptive metadata to authenticate the content. (Paynter et al., 2008)

4. Metadata

A digital object does not have any meaning to a human being unless the content is described with descriptive, structural and technical (or administrative) metadata. Preservation applications can only be successfully used when accompanied by supporting metadata.

Descriptive and preservation metadata assigned to digital research objects contains valuable information that can electronically be tracked by using software tools. Technical (or administrative) metadata that consists of two categories namely preservation and rights management metadata, are created to archive and sustain continuous access to data, from the origination of the digital asset to the storage of the final format of the object. This metadata aids in the long-term management of digital material and needs to be embedded in the planning processes. (Day, 2005)

The following technical (or administrative) metadata categories have been included in the research -

(1) Preservation

Preservation metadata contains archival information, which is needed for the longterm preservation of the object and the migration to other digital formats as software and hardware changes continuously

(2) Rights management

Technological mechanisms (Technical Protection Measures (TPM)), which restrict the usage of a digital object, can be embedded in the rights management metadata. Rights management metadata capture the permission of usage of an object and include the ownership, license information, restrictions on access, special permissions and methods of payment (if applicable).

The OAIS (www.oasis-open.org/home/index.php) (Open Archival Information System (ISO 14721:2002)) model introduces four new categories to the conventional standard metadata structure. These categories are grouped under the term Preservation Description Information (PDI). (Paradigm, 2005-7a)

(1) Reference Information

The reference information includes, and enumerate on specific identifiers which were assigned to the data, i.e. referencing such as ISBN number or Uniform Resource Name (URN)

(2) Provenance Information

The history of the content information (e.g., its origins, chain of custody, preservation actions and effects) is captured in the provenance information. This form of metadata helps to support a digital object's authenticity and integrity that is important for record-keeping and publication.

(3) Context Information

The context information index the relationship of the content to its environment (reason for creation, relationship to other data objects)

(4) Fixity Information

The fixity fields will document the authentication mechanisms, which in turn will ensure that the data is unaltered or show the extent of manipulation (e.g., checksum, digital signature). The checksum information can be used to implicate change in a stored file.

Completion of the above metadata has been applied to a variety of digital objects, during the course of the study. The purpose was to test the compliance of the metadata sets against the anticipated outcome for preservation of digital objects. Although the results of

the testing of the metadata were according to our expectations, a need for effective software for the management and retrieval of the objects through the metadata become clear.

The increasing amount of stored information impacts on the accessibility and preservation issues. The speed of retrieval and the ability to reproduce or retrieve information (within 24 hours in most cases) is a key factor for future deliverance of data. Time to retrieve a particular document can be reduced considerably by supplying sufficient metadata to index a document's content and history.

According to Rick Lawhorn the total worldwide digital archive capacity in the commercial and government sectors will grow to more than 27,000 petabytes or 27 exabytes by 2010. (Lawhorn, 2008)

5. Practical preservation applications

The term digital preservation refers to the preservation of materials that are digitally born and documents created with the use of imaging and recording technologies. Various views on the definition of preservation, and what is meant by preservation, exist. For the purpose of this study digital preservation is the preservation of digital materials for a period long enough in order for the object to survive the next generation of technology and software change without altering the original content of the source. (McKnight, 2003)

Repositories normally contain re-formatted copies of digital content for web display, i.e. MSWord converted to PDF-format. In most instances, the original digital object contains the archival value which should be stored as the master, ensuring that all changes can be tracked and interactions as well as relationships to other documents retained. Back-up copies of archival documents can be stored on trusted external hard drives and/or DVDs stored in controlled temperature conditions, inserted in acid-free pockets. However, the life span of this storage ware is dependent to the supportive technology. (Wellcome Library, 2008-).

5.1 Metadata added to document

Researchers, serious collectors of information and even users of information should know what metadata guidelines to use for capture, management, storage and/or preservation of digital objects. The future of digitised and born digital material, require significant thought and action. (Paradigm Project, 2005-07b)

Adopting good practice at the outset of a document will increase the longevity of the digital content. Additional to automated generated metadata, a table containing metadata of the specific document can be included in the document as well as stored separately as a "side-car".

5.2 Metadata examples

The following is an example of additional information that can be added to the body of a digital document which explain the format, and workflow/history of a digital document for preservation.

Fig. 3 represents an example of embedded metadata in a PDF-format document. This application in the software can be used for preservation metadata as it is searchable and will be migrated with the document format.

6. Management of digital content

6.1 Document management

The workflow of personal documentation can be managed through the use of a preadsheet with metadata fields and a Content Management System (e.g. Joomla) with capturing functions. The application of consistent file naming conventions (abbreviations of file content without any spaces between letters), standardisation on formats used, and metadata descriptions is recommended.

6.2 Metadata documentation

Compiling a spreadsheet for a collection i.e. images, will include all relevant preservation data, needed to recover a document. Description of the software used during the process, the workflow, editing dates, and specific manipulations done to the original images should be captured in such a document. This document needs to be stored separately from the original objects and preferably in more than one copy. Figure 4 shows an example of such a spreadsheet.

7. Search Functions

The search function plays an important role in the retrieval of electronic documents. On a personal computer with the Windows operating system, the MicroSoft Explorer search tool is commonly used for finding and retrieving of information. This search engine is not effective enough in specific hits. The Copernic tool was found to be valuable for desktop searching. This software searches a variety of categories, i.e. email, contact lists, files, music and the history of recently web visits with a structured display of the hits. (Figure 5).

Conclusions and Recommendations

The study helped us to benchmark the current situation regarding preservation in South Africa against the global awareness and actions on this topic. Negligence on format specifications and standardisation can cause huge data losses in the future and need further study towards a more simplified implementation of preservation strategies.

Storage and preservation of digital data need more attention in South Africa, and awareness towards preservation methods should be created amongst creators of digital content.

Training in the preservation of digital content and the actual delivery of plans and policies need to receive more attention in the corporate environment especially towards digital content stored on personal computers. Tools to help with digital preservation should be made more readily available and easier to install. Metadata, as a consistent, logical manner to keep documents accessible and usable over the long haul, need to be widely accepted and implemented, and the use thereof promoted to creators of digital objects.

References

Caplan, P. (2009,) *Understanding PREMIS (online)*, Library of Congress Network Development and MARC Standards Office, available at www.loc.gov/standards/premis/understanding-premis.pdf

Day, M. (2005), DDC/Digital curation manual instalment on metadata (online). HATII, University of Glasgow; University of Edinburgh; UKOLN, University of Bath; Council for the Central Laboratory of the Research Councils, available at http://www.dcc.ac.uk/resource/curation-manual/chapters/metadata

Jantz, R, Giarlo, M.J. (2005), "Digital preservation: architecture and technology for trusted digital repositories", Vol. 11, No. 6, *D-Lib Magazine (online)*, available at www.dlib.org/dlibjune05/jantz/06/jantz.html,

Lawhorn, R. (2008), Financial Security Tips: data protection essentials – a path to destruction, available at http://searchfinancialsecurity.techtarget.com/tip/0,289483,sid185_gci1294023,00.html

Library of Congress (2009), *Preservation: Preparing, protecting, preserving family treasures* (online), available at http://www.loc.gov/preserv/familytreasures/

McKnight, D. (2003), "DPI: The Digital Preservation Imperative (online)", Access 2003 Conference, Vancouver, BC, USA, 2 October 2003, available at http://access2003.lib.sfu.ca/presentations/mcknight/mcknight.ppt

Paradigm Project (2005-07a), Workbook on Digital Private Papers (online), available at http://www.paradigm.ac.uk/workbook

Paradigm Project (2005-07b), Guidance for creators of personal papers (online), available at http://www.paradigm.ac.uk/guidanceforcreators/guidance-for-creators-of-personal-digital-archives.pdf

Paynter, G., Joe, S., Lala, V., Lee, G. (2008), "A Year of Selective Web Archiving with the Web Curator at the National Library of New Zealand", *D-Lib Magazine* (online), vol. 14, no 5/6, available at http://www.dlib.org/dlib/may08/paynter/05paynter.html

University of London Computer Centre; UKOLN; JISC (2008), PoWR: The preservation of web resources handbook (online), available at http://jiscpowr.jiscinvolve.org/handbook

Wellcome Library (2008-), *Digital curation toolbox (online)*, available at http://library.wellcome.ac.uk/node289.html

About the authors

Ria Groenewald has been involved in several e-initiatives of the Department of Library Services, University of Pretoria since 2000. Ria is well trained in digitization and digital preservation, and has published a reward winning paper together with Amelia Breytenbach. She is the Digitization Coordinator of the UP LIS and is responsible for driving the digitization initiatives as well as for training and marketing thereof. She presented papers nationally and internationally and does undertake internal and external professional consultation where required. Ria is working on the digitization of the valuable collections housed at the University, and projects such as the digitization of the content of the Mapungubwe Museum, a Special Books project in collaboration with On-the-Dot, and the digitization of several Veterinary Science projects form part of her responsibility. She is a member of the Freedom Park Library and Knowledge Management Advisory Committee and also serves as a committee member of the South African Preservation and Conservation Association (SAPCON). Ria Groenewald is the corresponding author and can be contacted at: Ria.groenewald@up.ac.za

Amelia Breytenbach is an information and metadata specialist at the University Of Pretoria Department Of Library Services (Jotello F. Soga Library, Faculty of Veterinary Science), specializing in metadata and e-books. She holds a BBibl degree (1980) from the University of Pretoria. Her working experience covers a period of 29 years and includes cataloguing and classification of information resources with special interest in e-information resources, metadata applications and the building of information portals, blogs and websites. Key responsibilities in her current position regarding metadata are: metadata specialist for UPSpace (the institutional repository of the University of Pretoria), implementing the Dublin Core metadata tagging strategy, quality control of the metadata, internal and external metadata training, advice on all issues and decisions regarding metadata and monitoring and communicating new metadata developments and applications for the Dept. of Library Services. She is responsible for the coordination, development and metadata tagging of historical information resources for establishing the South African National Veterinary Repository (SANVR) for the Faculty of Veterinary Science. She is author and co-author of several papers presented at various conferences, workshops and training sessions and also an outstanding paper award winner at the Literati Network Awards for Excellence 2009.

Figures / Tables

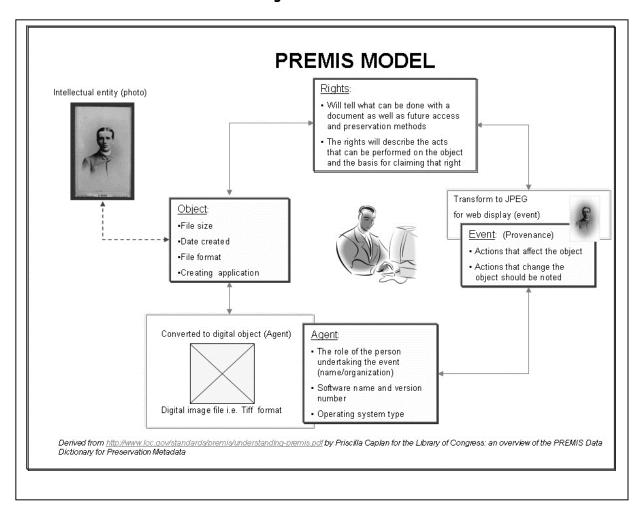


Fig.1: Explaining PREMIS

	1	
Document	The African Elephant: a digital collection of anatomical sketches as part of	
Title	the University of Pretoria's Institutional Repository - a case study	
Authors	Breytenbach, Amelia and Groenewald, Ria	
Description	Although several collections have been digitised and made available in the University of Pretoria's Institutional Repository, a pilot study has not been done to measure the project management and workflow. The collections available in the repository at the time of this project were all long-term projects. There was a need to identify a project small enough to conform to normal project management requirements to use as an example to establish the planning and workflow of future projects. This paper offers practical help to libraries starting with digitisation, it supplies valuable information for project management, planning of workflow and estimate time frames for completing a specific task in the digitization process.	
Date created	2007/09/28 -	
Rights	The authors. Document can be migrated for future usage.	
Туре	Article	
Access	Own use	Social network
		Repository
Format	MS Word 2003 (.doc)	
Format extent	3.62 MB (3,796,480 bytes)	
File name	2007_gro_bre	
Language	English	
Keywords	Digital storage ; Collections management ; University libraries ; Anatomical drawings ; South Africa	
Document History		
Version	Date	Comments
1	2007/09/28	Document created by authors
2	2007/11/20	Document edited by authors
3	2007/11/30	Final edit and submission to Journal

 $\label{table:tab$

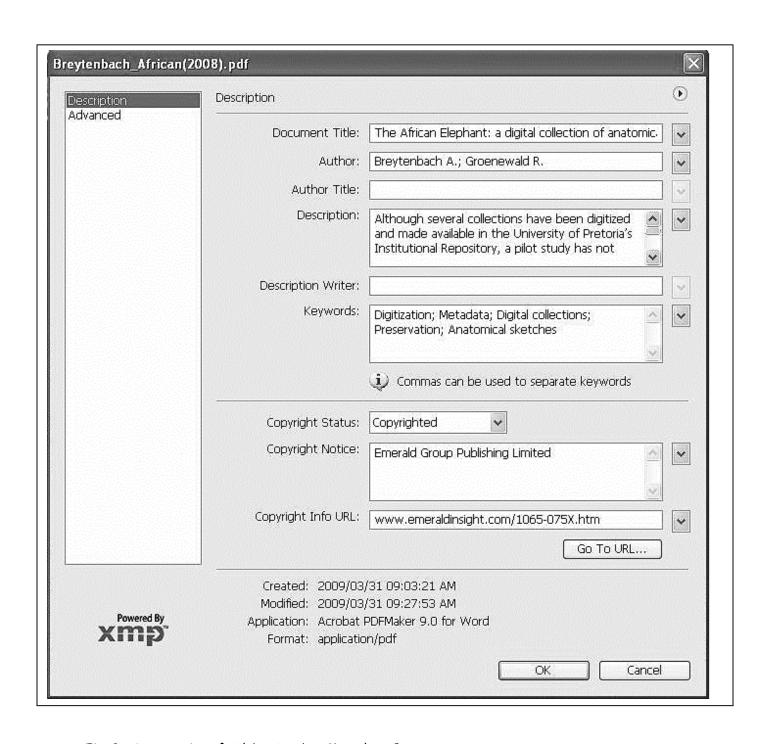


Fig. 2 : Screenprint of Adobe Acrobat Metadata Capture

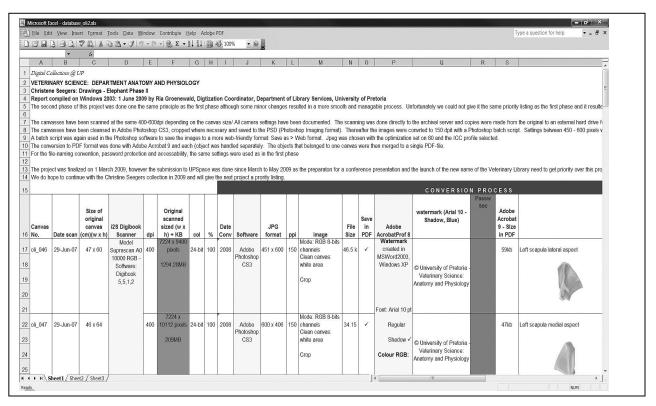


Fig.3: Spreadsheet with data about objects converted to digital format

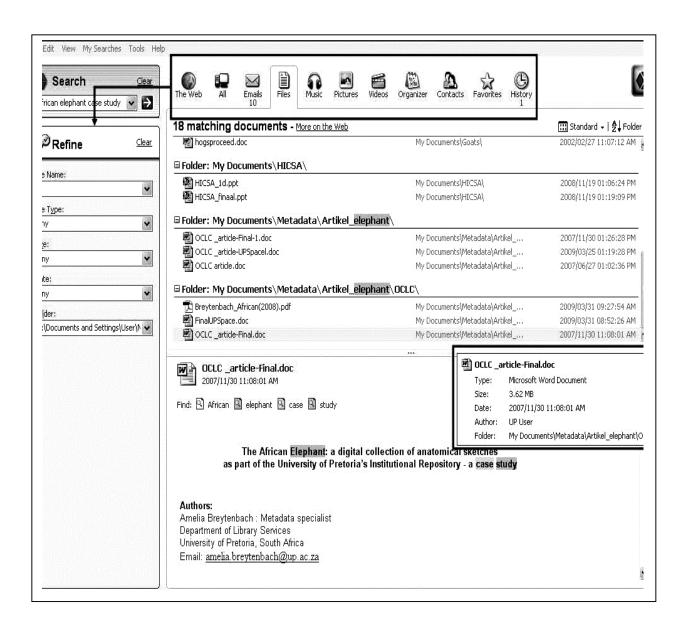


Fig. 4: Screenprint of Copernic search function