# Archives New Zealand Migration from Fedora Commons to the Rosetta Digital Preservation System

Jan Hutař
Digital Continuity
Government Digital Archive
Programme
Archives New Zealand
Te Tari Taiwhenua
Wellington, New Zealand
jan.hutar@dia.govt.nz

## ABSTRACT
This paper discusses the New Zealand Government Digital Archive Programme (GDAP) and its requirement for Archives New Zealand to move to a fully functional digital preservation system. It looks at the migration of digital content from Fedora Commons to Ex Libris' Rosetta Digital Preservation System focusing on what needed to be migrated, preparation of the migration, how it was performed and what tools were needed to support the work. We look at the verification of this process and conclude with an audit of the results and a description of the lessons learned during this process.

## Categories and Subject Descriptors
H.3.7 [**Digital Libraries**]: Systems Issues; J.1 [**Administrative Data Processing**]: Government

## General Terms
Management, Measurement, Verification.

## Keywords
Digital Archive, Migration, File Formats.

## 1. INTRODUCTION
Archives New Zealand, *Te Rua Mahara o te Kāwanatanga*, bound by the public records act 2005 **[1]** is the sole keeper of the memory of government in New Zealand. In 2010 it was announced by government that $12.6 million (NZD) would be made available to fund a Government Digital Archive (GDA), in order to improve management of the increasing number of digital records created by public sector agencies.

In support of this initiative, $9.7 million (NZD) was allocated to Archives New Zealand for the GDA across four years. $2.9 million (NZD) was allocated to the National Library of New Zealand for its equivalent programme, the National Digital Heritage Archive (NDHA). The goal was to work on the GDA project in co-operation with the library **[2]** by utilizing its existing systems. The NDHA programme started in 2005 and went live with its first digital preservation system in 2008.

Before GDAP, Archives New Zealand had the Digital Continuity Action Plan - endorsed by Cabinet in 2009. Building a robust digital archive system and processes is crucial to fulfilling statutory responsibilities for the long term preservation and accessibility of digital data from agencies. The GDA programme is one of the main outcomes of the Digital Continuity Action Plan.

As a part of the streamlining of government agency structures, The National Library of New Zealand and Archives New Zealand were formally incorporated into the Department of Internal Affairs (DIA) on February 2011. One of the consequences of this change was the decision that the GDA would leverage from previous government investment and research by sharing the existing digital preservation system of the NDHA – Rosetta **[3]**. Using one system required extending existing infrastructure, including hardware, architecture and the long-term preservation system settings, as well as the development of additional software capability. Another consideration is the development and understanding of organizational responsibilities and processes, as well as the creation of shared policies across both institutions.

The move to utilize the same systems developed by the NDHA required the migration of the content from Archives New Zealand's Interim Digital Archive (IDA), built on top of Fedora Commons, to Ex Libris' Rosetta digital preservation technology; this also required the integration of Archive New Zealand's "catalogue, collections management and public search" system - Archway. Migration of the IDA content is the first of the three releases planned as part of GDAP.

It is hoped that the new infrastructure will help Archives New Zealand to achieve five principle objectives **[4]**:

1. Protect important public sector digital information through change
2. Empower government, businesses, and communities to discover, access, understand, and reuse important public sector digital information
3. Foster digital continuity understanding with stakeholders
4. Streamline the transfer of information from public sector agencies to Archives New Zealand
5. Support the public sector to achieve the purposes of the Public Records Act 2005

## 2. REPOSITORY MIGRATION
### 2.1 Interim Digital Archive - Fedora Commons
Fedora Commons was selected and implemented in 2008 to provide Archives New Zealand with an Interim Digital Archive (IDA) after the establishment of the Digital Sustainability Programme. With longer term planning already underway for a programme to implement a complete digital preservation system,

IDA provided the organization with digital repository functionality that could potentially be replaced within 2-3 years. Active preservation of existing, archived digital materials was considered low priority for the IDA. Archives New Zealand identified several benefits of Fedora as a short-term solution for a digital repository including:

- Zero proprietary product costs and constraints
- Customizations being easier to make due to open source code base
- Fine grained security; support for up to one million digital objects
- The advantage that one other New Zealand government agency was using it - the State Services Commission, *Te Komihana O Ngā Tari Kāwanatanga* (SSC)

It was clear from the beginning that the Fedora based IDA would provide just the minimum functionality to support the business processes involved in accepting and managing a digital archive, that is, the ability to ingest data, manage archival objects and provide access to them via Archway. We knew it had limited functionality to support complex digital preservation. It was also necessary to build the IDA for the increasing number of materials being digitized for access. It was never used for storing data from physical carriers like floppy discs, CD/DVDs etc. which Archives New Zealand received from a handful of agencies. No digital transfer has ever been ingested into Fedora though it was one of the reasons for establishing it. The ability to accept digital transfers is one of the main deliverables of GDAP.

Fedora provided an adequate solution for an interim digital repository, but the technological infrastructure it was established on was limited. It was not a system that could be scaled to provide a 'whole-of-government' solution. The requirements of GDAP demanded more robust hardware and a logical digital preservation solution. So the decision was taken to align Archives New Zealand's technical approach for a digital repository with the well-established repository maintained by the National Library of New Zealand.

## 2.2 Rosetta

Rosetta is a long-term preservation system developed by Ex Libris. It may be considered an outcome of the NDHA programme, where initial requirements for such system began taking shape in 2005. This was originally in partnership with Endeavor Information Systems (Elsevier), later taken over by Ex Libris, who became the primary partner for developing the software package (after acquiring Endeavor in 2006 **[5]**). Rosetta has been used as a digital preservation system at the National Library of New Zealand since 2008, when its first version went live with the launch of the NDHA digital archive.

Presently, both institutions are using a shared implementation of Rosetta 3.1. There are currently 17 customers of this system around the world [Email communication with Nir Sherwinter (Ex Libris) on 4 April 2013].

## 2.3 Process

In preparation for the migration we needed to get details of the IDA content. The repository contained about 40TB of data at the initial stage of migration planning in late 2011. This became 48TB as data was ingested into the IDA during 2012, until the new digital repository was switched on in December 2012. The IDA mainly stored digitized documents from collections like the personnel records of First World War soldiers from the New Zealand Defence Force (NZDF); Westland maps; Land Information New Zealand (LINZ); the Treaty of Waitangi and other collections. Each collection had been appraised regarding the importance of the documents; the necessity of migration, that is, if they were already linked with Archway; and the difficulties expected in a migration. There was an Excel spreadsheet for each collection listing items' ID, title, description, collection ID, and the reason for migration or for leaving it out of the process.

Rosetta can ingest data in a certain shape and structure and with a certain metadata format. The Rosetta data model is based on the METS and PREMIS standards. Every Submission Information Package (SIP) ingested into the system has to be wrapped in METS with DNX metadata. DNX is Rosetta's proprietary metadata standard which can contain PREMIS-like metadata among technical metadata standards such as MIX.

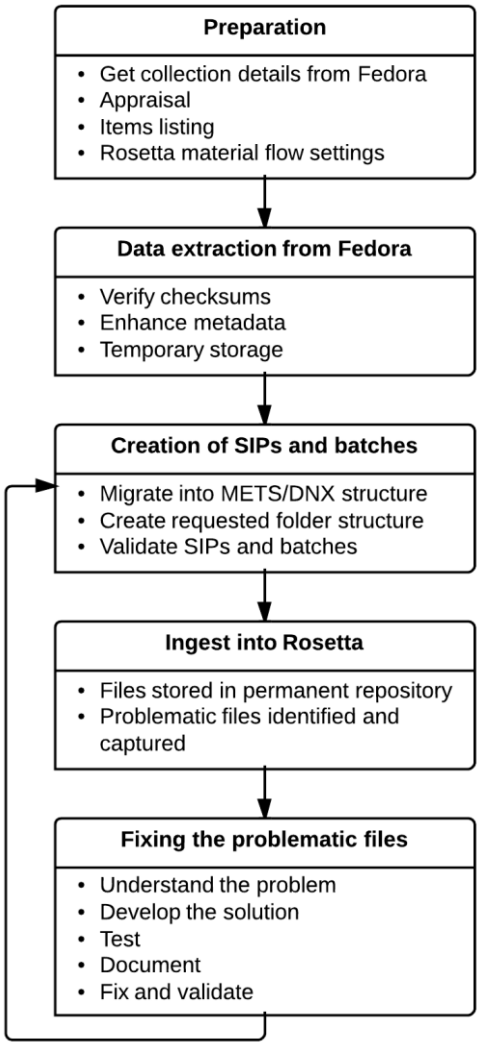
The First step of the data migration was extracting the digital objects and metadata from the Fedora repository. Objects were stored in Fedora with minimal metadata, because the descriptive metadata is stored in the Archway database, which is the archival description management system developed and used by Archives New Zealand. Key metadata for the migration was the Archway ID and checksum values. The Archway ID is used for linking between data in Rosetta and their description in Archway. If an MD5 or SHA-1 checksum was present for a file in Fedora, it was re-calculated after the file was extracted and compared against the stored value. Warnings were produced for:

- missing checksums,
- unsupported checksum types,
- failed checksum checks resulting in a failure for the item.

The next step was to migrate extracted objects and metadata into a new structure to comply with the Rosetta internal metadata standard and put them in the temporary storage location. SIPs can contain one or many items, but after testing it was decided to maintain a 1:1 ratio, that is, **one SIP for one item or record**. Each SIP has a METS structure and basic descriptive Dublin Core metadata (Title, Provenance, Series Number and Archway ID).

Rosetta will accept an item for deposit when packaged as part of a SIP. Generated SIPs were sorted by collection and arranged into suitably sized 'ingest batches' for overnight processing. Each batch of SIPs was prepared for pre-ingest, to check for zero-byte files, non-existent METS, and exceptionally long or improperly formatted filenames. We also checked for duplicate items, both within the same batch and against previously ingested ones. Finally, we triggered the ingest process in the Rosetta deposit module – see in the Figure 1 below.

**Figure 1: General workflow of the migration process**




## 2.4 Migration tools and Rosetta settings

For the data extraction from the IDA, its preparation and migration into the new SIP structure, a java based "Migration Tool" was developed. The tool recurses the IDA repository looking for all items with a published Archway ID and then takes the data from Fedora repository and puts it into a temporary location. The tool also invokes the process of creating the SIP in the structure Rosetta expects to receive, that is, with METS structure and additional administrative metadata (date of ingest, agent etc.).

Scripts were created to generate the batch for ingest, validate the batch and trigger the ingest itself via the Rosetta deposit API.

At the moment, we have a Producer entity (see below) in Rosetta based on the method of ingest. This means that for the automatic ingest of migrated data batches we created a new Producer called Archives New Zealand Digital Migration (ANZDM), in addition to the existing Archives New Zealand Internal Digitisation Programme (ANZIDP), which is used for all data coming from the Ingestor User Interface (UI) application. The Ingestor application is used by our archivists for ingesting digitized data on a daily basis. Each Producer in Rosetta has one or more agents that represent real actors from across Archives New Zealand, who are allowed to ingest if they have proper Rosetta and Ingestor roles and access rights. A Producer can be connected to one or more ingest material flows and could have different requirements for the metadata provided and file formats permitted for ingest etc.

Settings for the Fedora migration ingest material flow in Rosetta were the same as that for the normal ingest of digitized data via the Ingestor UI application. The steps of this ingest material flow include file format validation, virus check, risk assessment, structure validation, metadata extraction and access copy generation (for example TIFF to JPG). There were different material flows for different file collections; the only difference being the specifications used for access copies. For example, JPEG at 900x900 px resolution for NZDF A4 format documents compared to 3000x3000 px for maps. The goal was to avoid any unnecessary manual intervention. The only point where manual work is required is where files fail technical assessment, for example via DROID or JHOVE characterization, and end up in the Rosetta Technical Analyst Workbench for further investigation by a Digital Preservation Analyst.

The results, such as the number of ingested items and files, of each batch ingest could be confirmed by querying the Oracle database of Rosetta.

## 2.5 Result

There has been a total of 70 bulk ingests run over a period of 12 months - batch 001 on 7 February 2012 and batch 070 completed on 20 January 2013. The estimated total amount of data in the IDA was calculated at 48TB. On completion approximately 46TB (45,9TB) of data had been migrated, which represents 63,460 archival items (not files, item has 30 files on average). All of the items have been extracted, ingested into Rosetta and synchronized with the access portal Archway. The rest of the original 48TB were not migrated as they were not associated with any existing Archway item or temporary usage.

Items that were not successfully migrated into Rosetta, usually due to some technical problem associated with one or more file streams, were moved to a "quarantine" directory, which was kept on a NAS storage device. In total 453GB of data (0,1%), which equates to 468 items / SIPs have been quarantined.

Migration was scheduled across one year. On average, for each migration we ingested batches of 723GB in size. The biggest ingest was 1,2TB. Each batch consisted of anything between 10 to 5,000 items depending on the type of material being ingested. The limitation on size came from the expected ingest time required to process each batch. It was necessary for it to complete before 8am each day because we did not want this process running during normal working hours when it may impact on other processes. The average time taken for an ingest was seven hours. Performance was dependant on hardware configuration, volume of files and their file size. We were able to use our current hardware configuration with this number of batches and files; there was no requirement to complete ingests faster and therefore no need to upgrade the hardware for the Rosetta deposit module.

Another reason for restricting batch sizes was because of the time involved in auditing and reporting the results each day following the process. Problems that had occurred during a bulk ingest had to be resolved prior to preparation of the next ingest. It was only

possible to prepare the next batch on completion of the last, due to the risk of including undetected duplicate objects.

Our final audit was done via the Archway database, where all items are stored. It was compared with the original list of item IDs stored in IDA repository and then with the current Rosetta Oracle database of item IDs in our production environment. Put simply, if an IDA item has an associated Rosetta item ID, we can say that it has been synchronised with Archway via the Rosetta publishing process and therefore successfully migrated. We have identified only two duplicate items ingested during the entire operation.

# 3. FILE FORMATS

As mentioned, almost 0,5TB of data was identified as problematic and moved to quarantine outside of the Rosetta system. All of the issues related to problems with the bit-streams of files themselves, format identification, validation, or subsequent metadata extraction from these files. Tools like JHOVE generally will not validate files which do not conform to the specification of the format. Therefore in some cases no technical metadata is created, which is a major problem for us as we aim to create and keep as much technical and administrative metadata as possible. Also, we aim to have consistent metadata for similar file types. The IDA did not provide file format validation, identification and metadata extraction on ingest. No quality assurance on file formatting, validity, or structure was done, either for internal digitization outputs, or for digitized data received from external digitization companies. We have migrated only digitized documents and for that reason the file formats were limited only to TIFF files and PDF files.

## 3.1 General overview

The table below shows the list of issues we encountered ingesting digital objects from the IDA into Rosetta. All files were caught in the Rosetta Technical Analyst Workbench. In this environment the Technical Analyst can perform a technical assessment of each file and understand what is causing the issues, for example by looking at the JHOVE validation output, or messages from other tools. It is also possible to solve the issue; for example, in the case of multiple file format identification in DROID, choose the right identification; or in other cases download the file, investigate and fix the problems and upload the fixed file back into the workbench to be sent to the permanent archive after it is re-validated and relevant metadata extracted.

**Table 1: List of issues encountered during ingest into Rosetta**

| | Error Message | File Format | # of SIPs |
|---|---|---|---|
| 1 | Tag 305 out of sequence | TIFF | 197 |
| 2 | Tag 270 out of sequence, Tag 269 out of sequence | TIFF | 112 |
| 3 | Invalid ID in Trailer | PDF | 94 |
| 4 | Exception occurred during metadata extraction | PDF | 41 |
| 5 | Unknown field with tag 347 (0x15b) encountered. Invalid YCbCr subsampling. Cannot handle zero strip size missing an image filename | TIFF | 30 |
| 6 | Invalid DateTime separator | TIFF | 23 |
| 7 | Multiple formats found for file | TIFF | 4 |
| 8 | Checksum Error, Premature EOF | TIFF | 2 |
| 9 | Malformed dictionary: Vector must contain an even number of objects, but has 3 | PDF | 2 |
| 10 | Count mismatch for tag 36864; expecting 4, saw 0 | TIFF | 1 |
| 11 | Improperly nested array delimiters | PDF | 1 |
| 12 | Invalid character in hex string | PDF | 1 |
| 13 | Invalid page tree node | PDF | 1 |
| 14 | Invalid strip offset, JHOVE message: Invalid strip offset, Invalid DateTime separator: 2010/09/28 02:39:27 | TIFF | 1 |

In the process of migrating data from the IDA into Rosetta, we chose an approach more suitable for large amounts of data; we did not try to solve all issues in the Rosetta Technical Analyst Workbench, rather we moved all SIPs caught in the Technical Analyst Workbench to our own quarantine location. There, the digital objects were analysed, fixed in bulk with an agreed solution, and the whole SIP re-submitted into Rosetta. This allows Archives New Zealand to avoid too many individual issues sitting in the Technical Workbench to be resolved and to allow us to continue with the remainder of the migration process.

## 3.2 Issues

Error messages in Table 1 are mainly output by JHOVE. Issue 5 is from ImageMagick, which is used in Rosetta for creating JPG access copies from TIFF masters. If the creation of access copies is not done for a file, the whole SIP is routed to the Technical Analyst Workbench again. Below is short description of some of the issues we encountered.

The most frequent issue was related to the bad formatting of files, in particular TIFF files with their tags out of sequence. The error output *"Tag 305 Out of Sequence"* from JHOVE is a little misleading, in that the tag is not strictly out of sequence. The problem is that there are two TIFF 305 'Software' tags in the metadata, each containing a unique string value. Only one Software tag is permitted in the TIFF specification. This was a problem generated by one of scanners used by the digitization company that created these files.

A similar problem with "Tag 270 out of sequence, Tag 269 out of sequence" appeared in 112 SIP packages. This related to TIFF metadata, tag 270 ImageDescription and tag 269 DocumentName, which were populated accidentally by the digitization company.

The error Invalid ID trailer[1] in our PDF files was created because we merged two PDF files into one in our workflow for creating multipage PDF access copies. That is, PDF file containing all the scanned pages of a certain file was combined with a PDF cover

[1] ID entry is an array of two byte-strings constituting a file identifier for the file. File identifiers are defined by the optional ID entry in a PDF file's trailer dictionary **[6]**.

page with relevant information about the original file (Archway ID, Title etc.). The issue with the PDF trailer was caused by the PDF creation engine Multivalent used in our environment.

The ImageMagick error showing: "Unknown field with tag 347 (0x15b) encountered. Invalid YCbCr subsampling. Cannot handle zero strip size missing an image filename" was caused by the appearance of non-standard features of some TIFF files (JPEG compression in TIFF, PhotometricInterpretation TIFF baseline tag with YCbCr value etc). Again, this was different to other TIFF files from the same collection and was a processing error during the digitization and post processing. We also discovered that only libtiff v3.9.4 of ImageMagick had problems handling those TIFF files, previous and later versions of libtiff worked fine.

The final issue we should highlight was that of poorly formatted and thus invalid date time separators in TIFF baseline metadata tag 306 DateTime. The TIFF format standard **[7]** specifies that this value should be formatted as [YYYY:MM:DD HH:MM:SS], whereas all the ingested LINZ images had a "/" (forward slash) instead of a ":" colon, that is: [YYYY/MM/DD HH:MM:SS].

## 3.3 Common solutions

In order to complete the migration, we had to solve all of the issues and re-ingest the data into Rosetta. While it is possible to ignore errors, Archives New Zealand's policy is to deal with problems when they appear. Ignoring the problem would very likely cause other problems in the future, for example while trying to complete a preservation migration of poorly-formed file types, such as our TIFF examples, into a new preferred file format. We would not consider the list of issues serious and they are unlikely to cause problems rendering the file. Each file in the above examples could be rendered, but not always technical metadata was created by metadata extractors because the files were incorrectly formatted. If we were to ignore this it would mean that we have in our permanent archive some preservation master TIFF files with technical metadata and some without. This inconsistency could limit our ability to access and work with these files in future; for example, the ability to search based on metadata fields and then create sets of documents with certain features, or more importantly, to assess the risk linked to certain files and formats.

The majority of the issues were fixed with scripts developed in-house. These were sometimes very basic and might simply call relevant tools like ExifTool for changing the metadata. Each problem and its solution were thoroughly analyzed, tested and documented. The aim was to introduce as minimal a change as possible into the bit-stream of each of the relevant files. For analysis of erroneous files we used community standard tools such as JHOVE, DROID, NLNZ Metadata Extractor, FITS and basic hex editors.

Each issue has been thoroughly documented and that documentation has been saved in the Archives New Zealand EDRMS. EDRMS IDs of the documentation files were then added into the metadata of the corrected digital objects. The documentation consists of the problem description with links to the relevant file format documentation. There is a list of options for dealing with the problem and finally the decision about the preferred solution. Another part of documentation is about how the solution was tested. Custom scripts are also stored in the EDRMS. The idea behind this is that all changes to files have to be documented and referenced from the item metadata, so that future users can understand what was done and why.

All this is considered to be pre-conditioning and follows Archives New Zealand's Pre-conditioning Policy which was developed alongside the National Library of New Zealand. The Pre-conditioning Policy deals specifically with changes to digital content that has come within the control of the Archives or Library, but has not yet been ingested into the preservation system. It focuses on objects where there is a need to solve technical issues. The policy covers changes to content that do not result in both the original file and a copy being ingested. Pre-conditioning changes are made entirely on the original - they do not generate a new copy[2].

Conditions for pre-conditioning in the policy are that the nature of the change is completely reversible and not extensive; it cannot change the intellectual content and it must be documented. The preservation system must also store a provenance note. This note should remain as part of the file's preservation metadata throughout its existence. This is true for all changes of the digital objects and resulting metadata mentioned above.

The provenance note is automatically added into the metadata as an event of the preconditioning. It consists of a short description, the outcome and a reference to the documentation in the EDRMS. This process is part of each script used for solving the aforementioned issues.

# 4. CONCLUSION

Migrating 46TB of data is a big task. One would hope a minimal set of issues are likely to arise. To ensure a smooth migration, there are a couple of steps that need to be completed before the process begins. There has to be a plan, an analysis of current content, an ability to deal with issues and a mechanism for audit at the end. Handling issues as they occur and before ingest might prove to be the most time consuming part of the whole migration but ultimately makes the files more predictable to handle the next time around. In our case we had policies in place that helped speed up the decision making about what to do about different issues and these will continue to assist us in the future.

We have learned a lot from this migration. First of all, very few issues came from the migration itself. There was no lost or corrupted data. The main issue was the quality of the data, which had not been checked before that point. Data and file formats were not validated in the IDA solution. A key learning is that we now plan to do basic validation, with tools like JHOVE and DROID, as part of Archives New Zealand's internal process before accepting digitized data from external vendors.

Migration also helped us to understand the nature of the problems we will have to face once we start transfers of born-digital content from government agencies. Our approach to fix all the issues and keep as consistent an archive as possible might prove to be unrealistic while trying to cope with the flood of different types of digital objects from transfers.

We were also pleased to see that the Rosetta digital preservation system could easily cope with 1TB of data ingest in 6-8 hours

[2] If this is the case, its covered by the Preservation Action Policy and such a change must happen in the controlled environment of the preservation system.

within our current infrastructure. It was confirmation of our early expectations.

The last step of the migration was a final audit of the data in Rosetta and deletion of the Interim Digital Archive content - this was completed in July 2013.

## 5. ACKNOWLEDGMENTS

My thanks to colleagues Mike Ames, Ross Spencer, Tracie Almond, Matt Painter and GDAP manager Alison Fleming for being able to use internal documents which they have created.

## 6. REFERENCES

[1] Department of Internal Affairs of New Zealand. 2005. *Public Records Act*. Public Act 2005 No 40. http://www.legislation.govt.nz/act/public/2005/0040/latest/DLM345529.html

[2] New Zealand Government. 2010. *Announcement of Government Digital Archive* [online]. [cit. 20-04-2013]. http://www.beehive.govt.nz/speech/announcement-government-digital-archive

[3] Hutař, J. 2012. Assessing Digital Preservation Strategies. In *International Council on Archives Congress* (Brisbane, August 20 -24, 2012). http://www.ica2012.com/files/pdf/Full%20papers%20upload/ica12Final00155.pdf

[4] Archives New Zealand. 2010. *Government Digital Archive Programme* [online]. [cit. 20-04-2013]. http://archives.govt.nz/advice/government-digital-archive-programme

[5] Elsevier. 2006. *Francisco Partners to Acquire Endeavor Information Systems from Elsevier* [online]. [cit. 20-04-2013]. http://www.elsevier.com/about/press-releases/science-and-technology/francisco-partners-to-acquire-endeavor-information-systems-from-elsevier

[6] Adobe Systems Incorporated. 2000. *Adobe portable document format*. v 1.3. 2nd ed. p. 477. Addison-Wesley: Boston. ISBN 0-201-61588-6. http://wwwimages.adobe.com/www.adobe.com/content/dam/Adobe/en/devnet/pdf/pdfs/pdf_reference_archives/PDFReference13.pdf

[7] Aware systems. 2008. *TIFF Tag DateTime* [online]. [cit. 20-04-2013]. http://www.awaresystems.be/imaging/tiff/tifftags/datetime.html