Technical metadata, steppingstone to archiving?

Metadata is used by IT archiving as a starting point for moving to cheaper storage media.
I described the background of this process in my earlier article in Od 6 (September 2020) 'How does a computer archive'. Can metadata contribute to archiving the 80-85% of unstructured data without large investments in technology and using existing knowledge? To answer that question, we first need to know the background of technical metadata.

According to the Dutch National Archives, Metadata (also called metadata) are data that describe the characteristics of data. They are data about data. Examples of characteristics are the creator, the date of creation, the language used and the file format. Metadata describe not only the data itself, but also the context in which the data was created or received. And what has happened to the data since its creation or receipt'.
'Metadata provide information about other data, including a description of the data'. (Library Stanford),
'Metadaten erhöhen die Verständlichkeit von Daten für andere Wissenschaftlerinnen und Wissenschaftler, sie dienen der Reproduzierbarkeit von Forschung und der Auffindbarkeit von Daten'. (University of Hamburg),

Recurring in the definitions is that metadata is information about data.
According to Donna Burbank of DAMA (The Global Data Management Community), two types of Metadata can be distinguished:

1. Technical metadata
Describe the structure, format, and rules for storing data, in a file system and/or in a database system.

2. Business metadata
Describe the organizational definitions, rules, and context of data.

I limit myself to the background of technical metadata in the file system, because file data forms by far the largest part of the 80-85% of unstructured data. Technical metadata have been in use since the very beginning because they are indispensable for the proper functioning of a computer. Why are they indispensable?

Master File Table
A computer physically stores bits on a storage medium, such as a hard disk or a flash drive. This is regulated in the file system software, part of the operating software of the computer. Vendors have often created their own file system standards. The following explanation is based on Microsoft's, because by far the most unstructured documents of the 80-85% from the beginning of the 80's were stored using Microsoft filesystem software.

The Microsoft file system NTFS (New Technology File System) contains an entry called the Master File Table (MFT) (before 1993 File Allocation Table (FAT)). There is at least one entry in the master file table for each file on an NTFS file system volume, including the master file

table itself. All information about a file, including its size, time and date stamps, permissions, and data content, is stored in master file table entries. This technical metadata is called "attributes" by Microsoft. [Note 1]

It is through these master file table entries of technical metadata that the computer finds the documents (files) on the physical medium. The creation of this metadata by the file system is therefore essential.

Prioritize
We can conclude that technical metadata:
- Have been standardized for a long time within Microsoft environments
- Are always available because they are necessary for the operation of a computer
- Are always up to date through fully automated creation and mutation
- Form an index that can be consulted in Explorer

It appears that technical metadata have been linked to files according to a standard in FAT and NTFS since the early 1980s. This is interesting from an archiving point of view. Is there a basis for indexing without, for example, the direct availability of the applications belonging to the documents? Can we use this metadata to prioritize in the process towards final archiving and then eliminate the backlog manually or automatically in a structured way?

Grip
What would that look like in practice? Imagine a departmental share with many thousands of documents that has existed since the first networks in Novell's day. [Note 2] The documents have been physically transferred to new storage media every three to five years, in a logical sense they are still present in the same departmental share. The documents have not been archived according to archiving standards.

One approach might be to select by folder names (folders) that are likely to contain the most archival documents in the first instance. In the selected folders we then sort by creation date, which immediately gives us an impression of the oldest documents that are the first to qualify for archiving. If we first sort the documents by files of which the size, the creation date and the extension are the same, we immediately get a subset of documents that probably can be deduplicated quickly. This can then be done by the archivist on behalf of, or by the information owner himself, manually or automatically.

After deduplicating this selection, one can look for documents of which the file names match and of which a .doc and a .pdf copy exist. It is very likely that the outcome of this selection is relatively small and can then be archived, again by the archivist on behalf of, or by the information owner himself, manually or automatically.

This approach enables the archivist to provide an archiving service to the organization without in-depth knowledge of IT technology. By using technical metadata of documents, the archive department can increase its grip on the unstructured information in the organization and archive the 80-85% correctly with a structured approach.

[Note 1] The web archives. 2008, Microsoft, NTFS Metadata creation

[Note 2] Novell Netware was the most widely used networking software in the 80-90s to interconnect PCs and servers within organisations.

Published in OverheidsDocumentatie (Od) September 2020
Vakblad Overheiddocumentatie, Od, is the network platform for information professionals in the government and non-profit sector. With attention to developments in the field, best practices, collaboration, personal development, pitfalls, and dilemmas, we give a glimpse into the kitchen of the information professional at governments and the non-profit sector.

Author: Frank Jan Bertram
Frank Jan Bertram is a metadata researcher at HOTCOLDFROZENDATA
Frank Jan Bertram is the founder of HOTCOLDFROZENDATA, the new name for MDES, a data preservation specialist since 2001. MDES performed data measurements on organizations in the period from 2002 -2012. By measuring the total data of an organization, it turned out that the DMS database contained on average 10-15% of the total data.
He is also the initiator of the Ozzodata project.