organisation
information
archivierung

AWV AK 6.2

WARC ISO 28500

Frankfurt am Main

19. März 2012

Dr. Hubert Salm

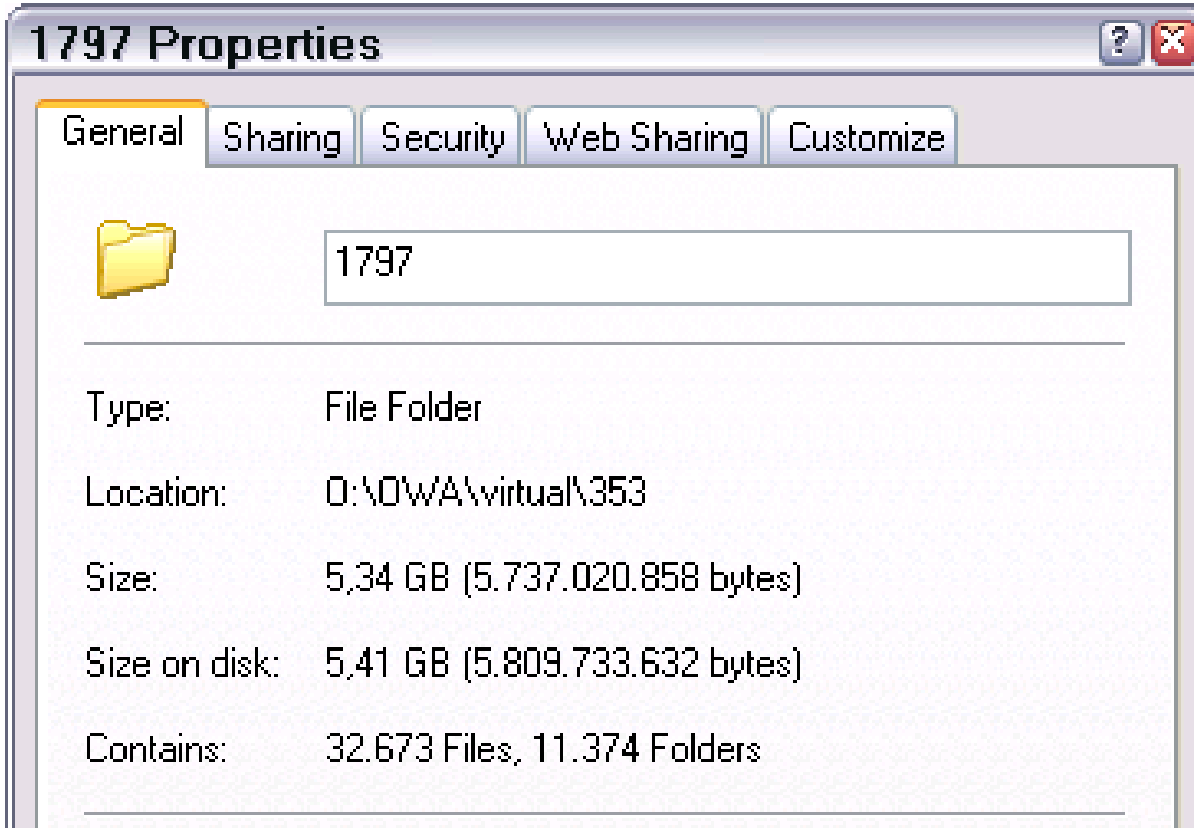
0. WARC - Literatur

From: Information and documentation — The WARC File Format
Date: 2008-XX-XX ISO 28500 ISO TC 46/SC 4/WG 12

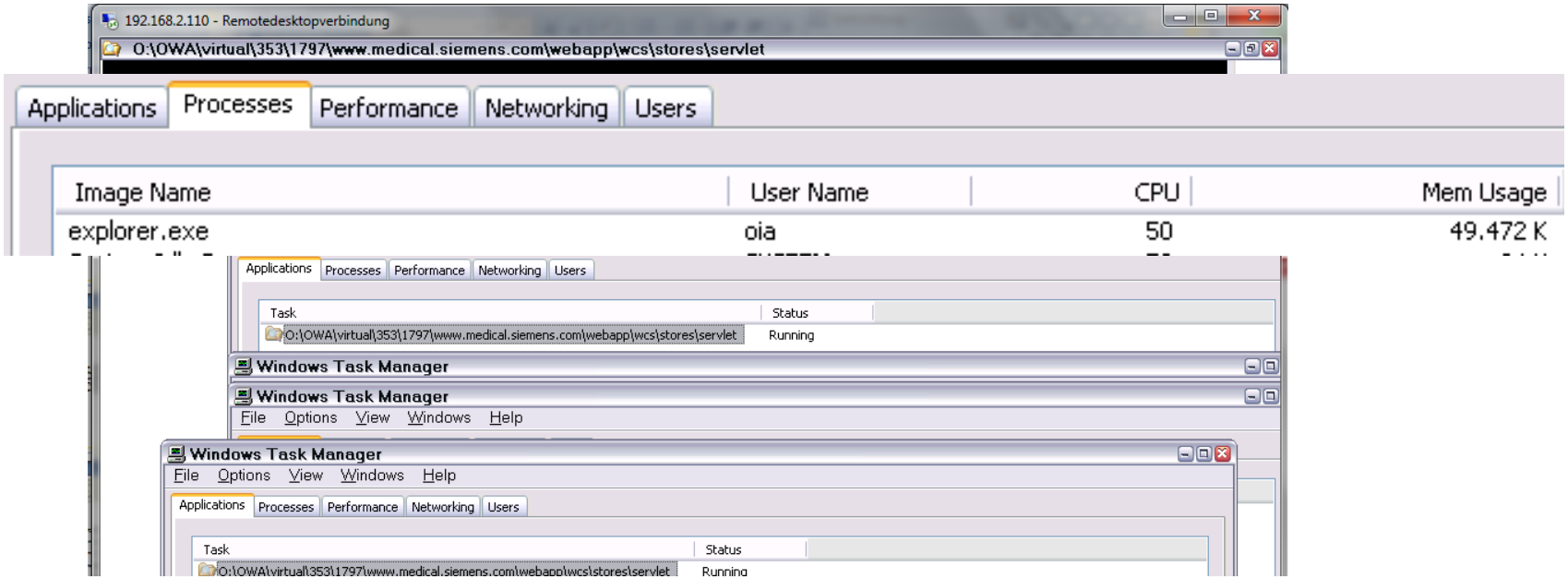
1. Von ARC (1998) nach WARC (2008)

At the same time, those same organizations have a rising need to archive large numbers of digital files not necessarily captured from the web (e.g., entire series of electronic journals, or data generated by environmental sensing equipment). A general requirement that appears to be emerging is for a container format that permits one file simply and safely to carry a very large number of constituent data objects for the purpose of storage, management, and exchange.

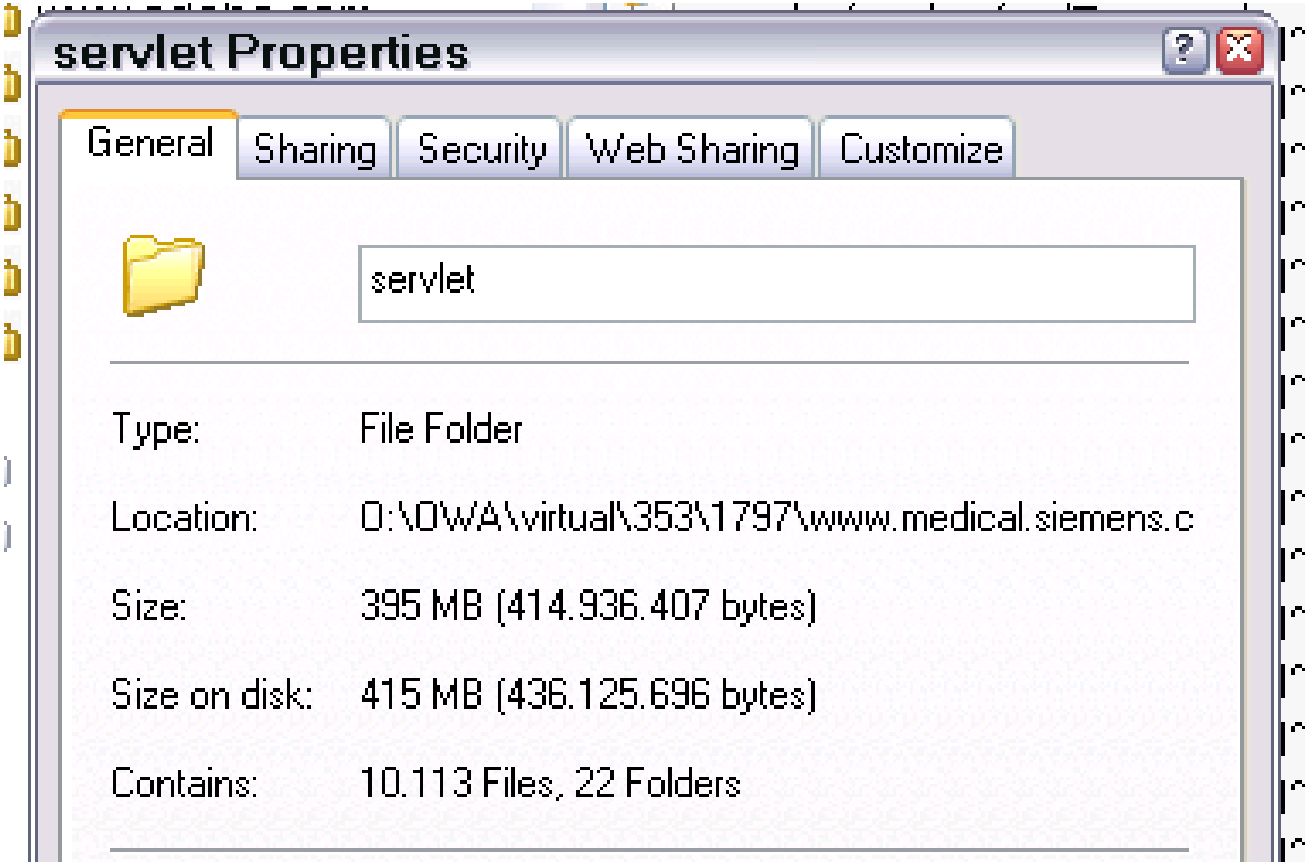
1.1 Wozu ARC? – Container im Archiv



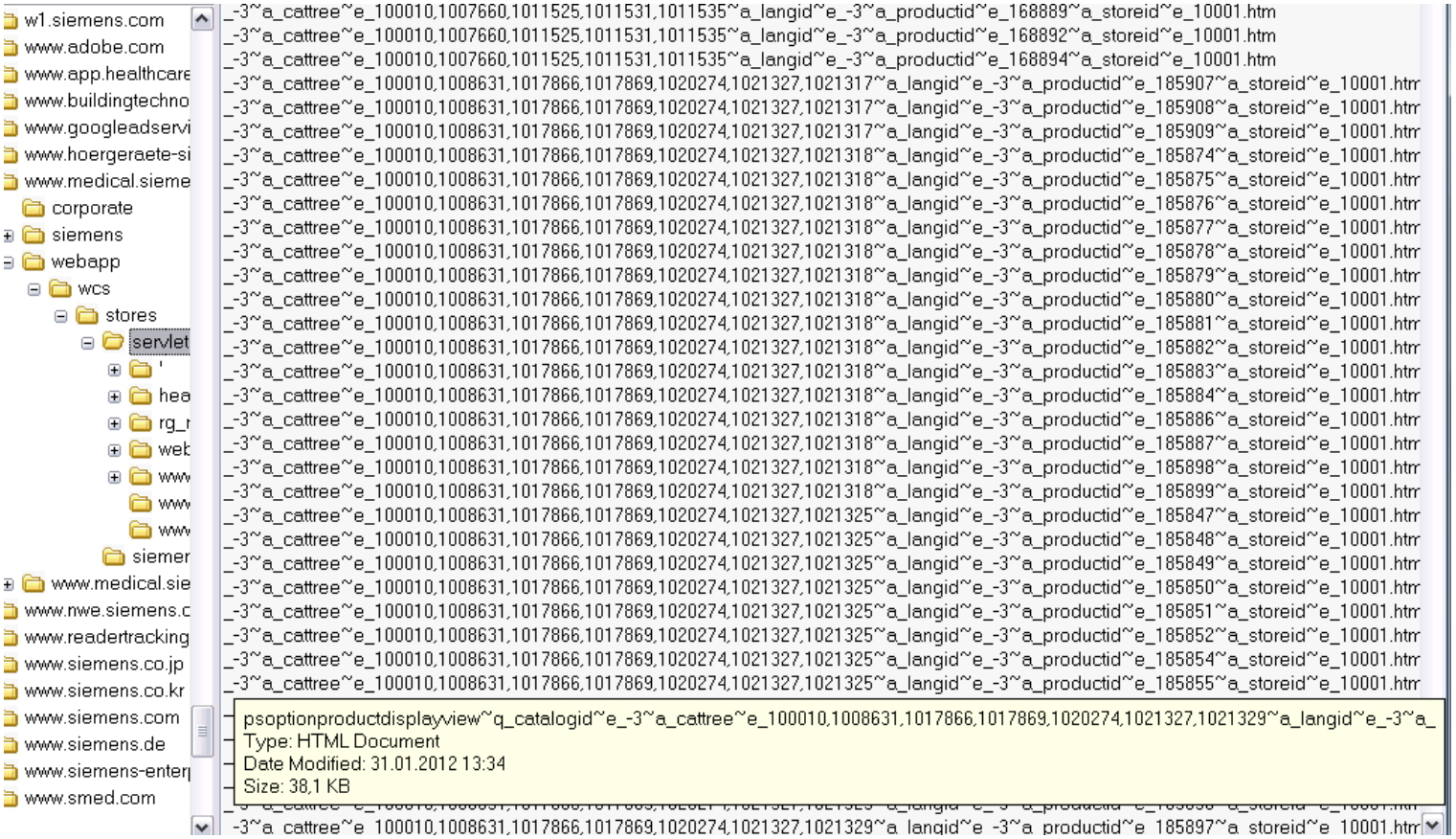
1.2 Wozu ARC?


192.168.2.110 - Remotedesktopverbindung
O:\OWA\virtual\353\1797\www.medical.siemens.com\webapp\wcs\stores\servlet
Applications
Processes
Performance
Networking
Users
Image Name
User Name
CPU
Mem Usage
explorer.exe
oia
50
49.472 K
Windows Task Manager
File Options View Windows Help
Task
Status
Running

1.3 Wozu ARC?


servlet Properties
General
Sharing
Security
Web Sharing
Customize
servlet
Type:
File Folder
Location:
O:\OWA\virtual\353\1797\www.medical.siemens.c
Size:
395 MB (414.936.407 bytes)
Size on disk:
415 MB (436.125.696 bytes)
Contains:
10.113 Files, 22 Folders

1.4 Wozu ARC


w1.siemens.com
www.adobe.com
www.app.healthcare
www.buildingtechno
www.googleadservi
www.hoergeraete-si
www.medical.sieme
corporate
siemens
webapp
wcs
stores
servlet
www.medical.sie
www.nwe.siemens.c
www.readertracking
www.siemens.co.jp
www.siemens.co.kr
www.siemens.com
www.siemens.de
www.siemens-enter
www.smed.com
psoptionproductdisplayview~q_catalogid~e_-3~a_cattree~e_100010,1008631,1017866,1017869,1020274,1021327,1021329~a_langid~e_-3~a_
Type: HTML Document
Date Modified: 31.01.2012 13:34
Size: 38,1 KB

2.0 Was ist WARC? - abstrakt

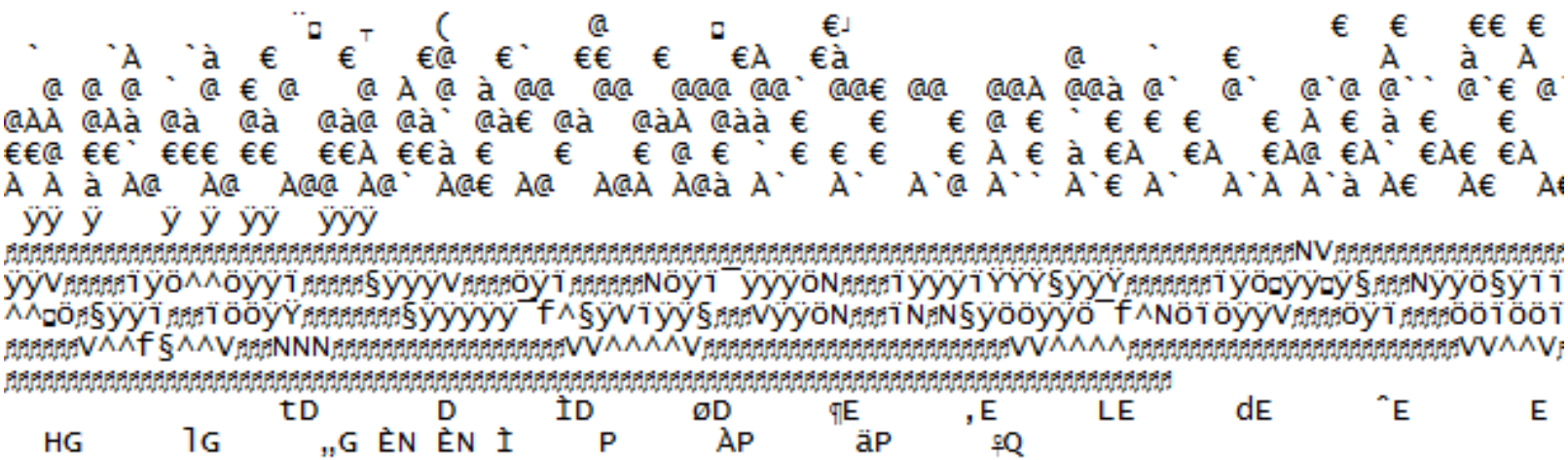
The WARC (Web ARChive) file format offers a convention for concatenating multiple resource records (data objects), each consisting of a set of simple text headers and an arbitrary data block into one long file.

2.1 Was ist WARC? - konkret

```
WARC/1.0
WARC-Type: request
WARC-Target-URI: http://www.oia-duesseldorf.de/images/favicon.ico
WARC-Date: 2012-02-22T11:13:03Z
WARC-Record-ID: <urn:uuid:04089cd0-7e99-4309-8b5c-d140c0aad192>
WARC-Warcinfo-ID: <urn:uuid:6655e77c-f0e1-4583-8e92-a8238fd0796e>
WARC-Concurrent-To: <urn:uuid:a4db77f7-bd1f-4fdb-8aec-40ff849888e1>
Content-Type: application/http;msgtype=request
Content-Length: 265

GET /images/favicon.ico HTTP/1.1
User-Agent: Mozilla/5.0 (compatible; MSIE 9.0; Windows; Windows NT 6.1; WOW64; .NET CLR 1.0.3705; Trident/5.0) Gecko/20091221 Firefox/3.5.7 oia.OWA
Host: www.oia-duesseldorf.de
Cache-Control: no-store,no-cache
Pragma: no-cache


WARC/1.0
WARC-Type: response
WARC-Target-URI: http://www.oia-duesseldorf.de/images/favicon.ico
WARC-Date: 2012-02-22T11:13:03Z
WARC-Record-ID: <urn:uuid:a4db77f7-bd1f-4fdb-8aec-40ff849888e1>
WARC-Warcinfo-ID: <urn:uuid:6655e77c-f0e1-4583-8e92-a8238fd0796e>
WARC-Payload-Digest: sha1:ZNXKKFJA76SLL6KNKFXPCBX75PI6BWMD
Content-Type: application/http;msgtype=response
Content-Length: 2784

HTTP/1.1 200 OK
MicrosoftOfficeWebServer: 5.0_Pub
Accept-Ranges: bytes
Content-Length: 2494
Content-Type: image/x-icon
Date: Wed, 22 Feb 2012 11:13:02 GMT
ETag: "8062f6338717cb1:c477c"
Last-Modified: Tue, 29 Jun 2010 12:32:57 GMT
Server: Microsoft-IIS/6.0
X-Powered-By: ASP.NET
```

2.2 Was ist WARC? – konkret, Detail

```
HTTP/1.1 200 OK
MicrosoftOfficeWebServer: 5.0_Pub
Accept-Ranges: bytes
Content-Length: 2494
Content-Type: image/x-icon
Date: Wed, 22 Feb 2012 11:13:02 GMT
ETag: "8062f6338717cb1:c477c"
Last-Modified: Tue, 29 Jun 2010 12:32:57 GMT
Server: Microsoft-IIS/6.0
X-Powered-By: ASP.NET
```



2.3 Warum WARC?

The WARC format is expected to be a standard way to structure, manage and store billions of resources collected from the web and elsewhere. It will be used to build applications for harvesting (such as the open source Heritrix web crawler), managing, accessing, and exchanging content. The way WARC files will be created and resources will be stored and rendered will depend on software and applications implementations.

3 WARC – weitere Eigenschaften

Besides the primary content recorded in ARCs, the extended WARC format accommodates

- related secondary content, such as assigned metadata,
- abbreviated duplicate detection events,
- later-date transformations, and
- segmentation of large resources.

3.1 WARC – Namenskonventionen

It is helpful to use practices within an institution that make it unlikely or impossible to duplicate aggregate WARC file names. The convention used inside the Internet Archive with ARC files is to name files according to the following pattern: Prefix-Timestamp-Serial-Crawlhost.warc.gz.

Prefix is an abbreviation usually reflective of the project or crawl that created this file.

Timestamp is a 14-digit GMT timestamp indicating the time the file was initially begun.

Serial is an increasing serial-number within the process creating the files, often (but not necessarily) unique with regard to the Prefix.

Crawlhost is the domain name or IP address of the machine creating the file.

BEISPIEL aus der OWA Implementation:

OWA-oia-00375-20120222120955-01842-00000-OIA-OWA.warc.gz

3.2 WARC – interne Struktur (Beispiel)

3.1.4 WARC record header

Beginning of a WARC record, consisting of one first line declaring the record to be in the WARC format with a given version number, followed by lines of named fields up to a blank line.

Implementationsbeispiel:

WARC/1.0
WARC-Type: warcinfo
WARC-Filename: OWA-oia-20120222120955-00375-01842-00000-OIA-OWA.warc.gz
WARC-Date: 2012-02-22T11:12:06Z
WARC-Record-ID: <urn:uuid:6655e77c-f0e1-4583-8e92-a8238fd0796e>
Content-Type: application/warc-fields
Content-Length: 11822

3.3 WARC Kompression (optional)

The WARC format defines no internal compression. Whether and how WARC files should be compressed is an external decision. However, experience with the precursor ARC format at the Internet Archive has demonstrated that applying simple standard compression can result in significant storage savings, while preserving random access to individual records. For this purpose, the GZIP format with customary "deflate" compression is recommended, as defined in [RFC1950], [RFC1951], and [RFC1952]. Freely available source code implementing this format is available, and the technique is free of patent encumbrances. The GZIP format is also widely used and supported across many free and commercial software packages and operating systems.
This section documents recommended, but optional, practices for compressing WARC files with GZIP.

4.1 WARC record type „warcinfo“

There are eight WARC record types: 'warcinfo', 'response', 'resource', 'request', 'metadata', 'revisit', 'conversion', and 'continuation'.

4.2 Crawler Metadata - Implementierung

5.8 WARC-Block-Digest

An optional parameter indicating the algorithm name and calculated value of a digest applied to the full
block of the record.

WARC-Block-Digest = "WARC-Block-Digest" ":" labelled-digest

labelled-digest = algorithm ":" digest-value

algorithm = token

digest-value = token

An example is a SHA-1 labelled Base32 ([RFC3548]) value:

WARC-Block-Digest: sha1:AB2CD3EF4GH5IJ6KL7MN8OP

Crawler Metadata - Implementierung

Implementierung, Beispiel:

software: oGet/10.2.0 http://www.oia-duesseldorf.de
ip: 192.168.2.110
hostname: OIA-OWA
format: WARC File Format 1.0
conformsTo: ISO 28500:2009 (http://bibnum.bnf.fr/WARC/WARC_ISO_28500_version1_latestdraft.pdf)
http-header-user-agent: Mozilla/5.0 (compatible; MSIE 9.0; Windows; Windows NT 6.1; WOW64; .NET CLR 1.0.3705; Trident/5.0) Gecko/20091221 Firefox/3.5.7 oia.OWA
crawler-Metadata:

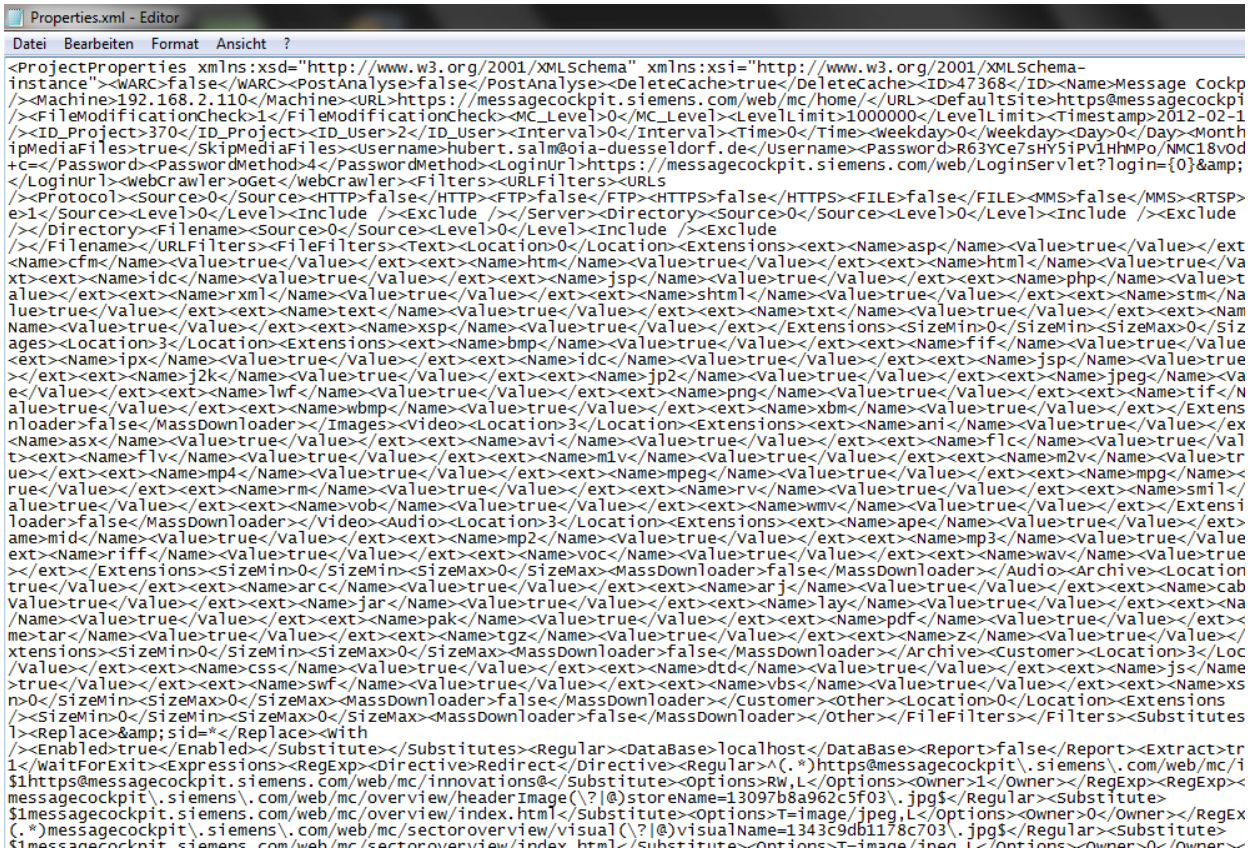
HRIHE33KMVRXIUDSN5YGK4TUNFSXGIDYNVWG44Z2PBZWIPJCNB2HI4B2F4XXO53XFZ3TGLTPOJTS6MRQGAYS6WCNJRJWG2DFNVQSEIDYNVWG44Z2PBZWSPJCNB2HI4B2F4XXO53XFZ3TGLTPOJTS6MRQGAYS6WCNJRJWG2DFNVQS22LOO
N2GC3TDMURD4PCXIFJEGPTUOJ2WKPBPK5AVEQZ6HRIG643UIFXGC3DZONST4ZTBNRZWKPBPKBXXG5CBNZQWY6LTMU7DYRDFNRSXIZKDMFRWQZJ6ORZHKZJ4F5CGK3DFORSUGYLDNBST4PCJIQ7DINZTHE4DYL2JIQ7DYTTBNVST433JME
6C6TTBNVST4PCDNB

Metadaten, strukturierte Übersicht

Steuerungsdaten (Authentizität)

Feld	Wert (Beispiel)	Quelle	Funktion	Bemerkung	Bem.
Erfassungsanlass	Umorganisation; Wahl	Manuell		Anlass, Intervall	wird benötigt; Intervall: automatisch
Transfertyp	Import oia-Spiegelungsauftrag	Manuell		Woher kommen die Daten, z.B. aus Archive.org, sonstiger Import	wird benötigt
Methode				Spiegelung / FTP	nb
Erfassungsdatum		System		Von bis gelaufen	ja

OWA, implementierte Metadaten WARC + CO


Properties.xml - Editor

Datei Bearbeiten Format Ansicht ?

<ProjectProperties xmlns:xsd="http://www.w3.org/2001/XMLSchema" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"><WARC>false</WARC><PostAnalyse>false</PostAnalyse><DeleteCache>true</DeleteCache><ID>47368</ID><Name>Message Cockp
/><Machine>192.168.2.110</Machine><URL>https://messagecockpit.siemens.com/web/mc/home/</URL><DefaultSite>https@messagecockpi
/><FileModificationCheck>1</FileModificationCheck><MC_Level>0</MC_Level><LevelLimit>1000000</LevelLimit><Timestamp>2012-02-1
/><ID_Project>370</ID_Project><ID_User>2</ID_User><Interval>0</Interval><Time>0</Time><Weekday>0</Weekday><Day>0</Day><Month
ipMediaFiles>true</SkipMediaFiles><Username>hubert.salm@oia-duesseldorf.de</Username><Password>R63YCe7sHY5iPV1HhMPo/NMC18vOd
+c=</Password><PasswordMethod>4</PasswordMethod><LoginUrl>https://messagecockpit.siemens.com/web/LoginServlet?login={0}&
</LoginUrl><WebCrawler>oGet</WebCrawler><Filters><URLFilters><URLs
/><Protocol><Source>0</Source><HTTP>false</HTTP><FTP>false</FTP><HTTPS>false</HTTPS><FILE>false</FILE><MMS>false</MMS><RTSP>
e>1</Source><Level>0</Level><Include /><Exclude /></Server><Directory><Source>0</Source><Level>0</Level><Include /><Exclude
/></Directory><Filename><Source>0</Source><Level>0</Level><Include /><Exclude
/></Filename></URLFilters><FileFilters><Text><Location>0</Location><Extensions><ext><Name>asp</Name><Value>true</Value></ext
<Name>cfm</Name><Value>true</Value></ext><ext><Name>htm</Name><Value>true</Value></ext><ext><Name>html</Name><Value>true</Va
xt><ext><Name>idc</Name><Value>true</Value></ext><ext><Name>jsp</Name><Value>true</Value></ext><ext><Name>php</Name><Value>t
alue></ext><ext><Name>rxml</Name><Value>true</Value></ext><ext><Name>shtml</Name><Value>true</Value></ext><ext><Name>stm</Na
lue>true</Value></ext><ext><Name>text</Name><Value>true</Value></ext><ext><Name>txt</Name><Value>true</Value></ext><ext><Nam
Name><Value>true</Value></ext><ext><Name>xsp</Name><Value>true</Value></ext></Extensions><SizeMin>0</SizeMin><SizeMax>0</Siz
ages><Location>3</Location><Extensions><ext><Name>bmp</Name><Value>true</Value></ext><ext><Name>fif</Name><Value>true</Value
<ext><Name>ipx</Name><Value>true</Value></ext><ext><Name>idc</Name><Value>true</Value></ext><ext><Name>jsp</Name><Value>true
></ext><ext><Name>j2k</Name><Value>true</Value></ext><ext><Name>jp2</Name><Value>true</Value></ext><ext><Name>jpeg</Name><Va
e</Value></ext><ext><Name>lwf</Name><Value>true</Value></ext><ext><Name>png</Name><Value>true</Value></ext><ext><Name>tif</N
alue>true</Value></ext><ext><Name>wbmp</Name><Value>true</Value></ext><ext><Name>xbm</Name><Value>true</Value></ext></Extens
nloader>false</MassDownloader></Images><Video><Location>3</Location><Extensions><ext><Name>ani</Name><Value>true</Value></ex
<Name>asx</Name><Value>true</Value></ext><ext><Name>avi</Name><Value>true</Value></ext><ext><Name>flc</Name><Value>true</Val
t><ext><Name>flv</Name><Value>true</Value></ext><ext><Name>m1v</Name><Value>true</Value></ext><ext><Name>m2v</Name><Value>tr
ue></ext><ext><Name>mp4</Name><Value>true</Value></ext><ext><Name>mpeg</Name><Value>true</Value></ext><ext><Name>mpg</Name><
rue</Value></ext><ext><Name>rm</Name><Value>true</Value></ext><ext><Name>rv</Name><Value>true</Value></ext><ext><Name>smil</
alue>true</Value></ext><ext><Name>vob</Name><Value>true</Value></ext><ext><Name>wmv</Name><Value>true</Value></ext></Extensi
loader>false</MassDownloader></Video><Audio><Location>3</Location><Extensions><ext><Name>ape</Name><Value>true</Value></ext>
ame>mid</Name><Value>true</Value></ext><ext><Name>mp2</Name><Value>true</Value></ext><ext><Name>mp3</Name><Value>true</Value
ext><Name>riff</Name><Value>true</Value></ext><ext><Name>voc</Name><Value>true</Value></ext><ext><Name>wav</Name><Value>true
></ext></Extensions><SizeMin>0</SizeMin><SizeMax>0</SizeMax><MassDownloader>false</MassDownloader></Audio><Archive><Location
true</Value></ext><ext><Name>arc</Name><Value>true</Value></ext><ext><Name>arj</Name><Value>true</Value></ext><ext><Name>cab
Value>true</Value></ext><ext><Name>jar</Name><Value>true</Value></ext><ext><Name>lay</Name><Value>true</Value></ext><ext><Na
/Name><Value>true</Value></ext><ext><Name>pak</Name><Value>true</Value></ext><ext><Name>pdf</Name><Value>true</Value></ext><
me>tar</Name><Value>true</Value></ext><ext><Name>tgz</Name><Value>true</Value></ext><ext><Name>z</Name><Value>true</Value></
xtensions><SizeMin>0</SizeMin><SizeMax>0</SizeMax><MassDownloader>false</MassDownloader></Archive><Customer><Location>3</Loc
/Value></ext><ext><Name>css</Name><Value>true</Value></ext><ext><Name>dtd</Name><Value>true</Value></ext><ext><Name>js</Name
>true</Value></ext><ext><Name>swf</Name><Value>true</Value></ext><ext><Name>vbs</Name><Value>true</Value></ext><ext><Name>xs
n>0</SizeMin><SizeMax>0</SizeMax><MassDownloader>false</MassDownloader></Customer><Other><Location>0</Location><Extensions
/><SizeMin>0</SizeMin><SizeMax>0</SizeMax><MassDownloader>false</MassDownloader></Other></FileFilters></Filters><Substitutes
l><Replace>&sid=*</Replace><With
/><Enabled>true</Enabled></Substitute></Substitutes><Regular><DataBase>localhost</DataBase><Report>false</Report><Extract>tr
1</WaitForExit><Expressions><RegExp><Directive>Redirect</Directive><Regular>^(.*)https@messagecockpit\.siemens\.com/web/mc/i
$1https@messagecockpit.siemens.com/web/mc/innovations@</Substitute><Options>RW,L</Options><Owner>1</Owner></RegExp><RegExp><
messagecockpit\.siemens\.com/web/mc/overview/headerImage(\?|@)storeName=13097b8a962c5f03\.jpg$</Regular><Substitute>
$1messagecockpit.siemens.com/web/mc/overview/index.html</Substitute><Options>T=image/jpeg,L</Options><Owner>0</Owner></RegEx
(.*)messagecockpit\.siemens\.com/web/mc/sectoroverview/visual(\?|@)visualName=1343c9db1178c703\.jpg$</Regular><Substitute>

OWA, implementierte Metadaten WARC + CO

```
<?xml version="1.0"?>
- <ProjectProperties xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:xs
    <WARC>false</WARC>
    <PostAnalyse>false</PostAnalyse>
    <DeleteCache>true</DeleteCache>
    <ID>47368</ID>
    <Name>Message Cockpit</Name>
    <CharCode/>
    <Machine>192.168.2.110</Machine>
    <URL>https://messagecockpit.siemens.com/web/mc/home/ </URL>
    <DefaultSite>https@messagecockpit.siemens.com/web/index.html</DefaultSite>
    <Macros/>
    <FileModificationCheck>1</FileModificationCheck>
    <MC_Level>0</MC_Level>
    <LevelLimit>1000000</LevelLimit>
    <Timestamp>2012-02-15T15:42:32.43934+01:00</Timestamp>
    <Description/>
    <ID_Project>370</ID_Project>
    <ID_User>2</ID_User>
    <Interval>0</Interval>
    <Time>0</Time>
```

Das Metadatenprojekt (AWV AK 6.2)

AWV AK 5.2

Hersteller, z.B. oia

Implementierung, z.B. oia OWA

WARC Container