

Practical Office Automation

Hacking the OpenOffice.org File Format

Jacob Sparre Andersen

sparre@crs4.it



GULCh

Gruppo Utenti Linux Cagliari h...?

Overture

- **Subject:**

This talk is about extracting and using meta-data from the OpenOffice.org/OpenDocument file format.

- **Audience:**

System administrators, system programmers and information system decision makers.

I will talk about what you can (tell your programmers to) do, if your documents are in an open format. Once I have told you what you *can* do, I will give you some examples of *how* to do it with standard Linux tools.

Feel free to ask questions at *any* time during the talk.

Overview (1)

- *What you can do, if your documents are in an open format.*
- A look into an OpenOffice.org file.
- Indexing OpenOffice.org documents.
- Preventing document histories from leaking out through your firewall.

Open file formats

The minimum requirements for an open standard are that the document format is completely described in publicly accessible documents, [...] and that the document format may be implemented in programs without restrictions, royalty-free, and with no legal bindings.

<http://europa.eu.int/idabc/servlets/Doc?id=17982>

Benefits from using open file formats

- Not tied to a single software provider.
- Lower price on off-the-shelf software.
- Freedom to (make your programmers) implement special in-house tools.
- It is more likely that you can find Open Source programs which already solve your problems.

Ideas for special in-house tools

Since the OpenOffice.org/OpenDocument file format is open, you are free to (make your programmers) write special tools for manipulating the files:

- Extracting titles and keywords for automated document indices.
- Blocking documents containing their editing history from exiting through the corporate firewall.
- Warning authors about lacking project codes in documents.
- ...

Only your imagination (and your ability to explain your ideas) sets limits.

Overview (2)

- What you can do, if your documents are in an open format.
- *A look into an OpenOffice.org file.*
- Indexing OpenOffice.org documents.
- Preventing document histories from leaking out through your firewall.

Looking into an OpenOffice.org file (1)

```
% unzip -l skriv-og-slet.sxw
Length      Date       Time       Name
-----
      30    11-14-05   11:40     mimetype
    1958    11-14-05   11:40     content.xml
    5979    11-14-05   11:40     styles.xml
    1282    11-14-05   11:40     meta.xml
    6280    11-14-05   11:40     settings.xml
     752    11-14-05   11:40     META-INF/manifest.xml
-----
    16281                                6 files
```


Looking into an OpenOffice.org file (2)

```
% unzip -ap skriv-og-slet.sxw meta.xml \  
> | sed 's/></>\n</g'  
<?xml version="1.0" encoding="UTF-8"?>  
<!DOCTYPE office:document-meta PUBLIC "-//OpenOf  
<office:document-meta xmlns:office="http://openc  
<office:meta>  
<meta:generator>OpenOffice.org 1.1.4 (Unix)</met  
<!--645 (Build:8824)-->  
<dc:title>Writes and deletions</dc:title>  
<meta:creation-date>2005-11-14T12:31:10</meta:cr  
<dc:date>2005-11-14T12:40:48</dc:date>  
<meta:keywords>
```

Looking into an OpenOffice.org file (3)

```
% unzip -ap skriv-og-slet.sxw meta.xml \  
> | sed 's/></>\n</g' \  
> | grep '<meta:keyword>'  
<meta:keyword>00o</meta:keyword>  
<meta:keyword>file format</meta:keyword>  
<meta:keyword>demonstration</meta:keyword>  
<meta:keyword>changes</meta:keyword>  
%
```

Looking into an OpenOffice.org file (4)

```
% unzip -ap skriv-og-slet.sxw meta.xml \  
> | sed 's/></>\n</g' \  
> | grep '<dc:title>'  
<dc:title>Writes and deletions</dc:title>  
%
```

Looking into an OpenOffice.org file (5)

```
% unzip -ap skriv-og-slet.sxw content.xml \  
> | sed 's/></>\n</g' \  
> | grep '<text:tracked-changes>'  
<text:tracked-changes>  
%
```

Overview (3)

- What you can do, if your documents are in an open format.
- A look into an OpenOffice.org file.
- *Indexing OpenOffice.org documents.*
- Preventing document histories from leaking out through your firewall.

Indexing OpenOffice.org documents

Practical demonstration of indexing of OpenOffice.org documents.

Overview (4)

- What you can do, if your documents are in an open format.
- A look into an OpenOffice.org file.
- Indexing OpenOffice.org documents.
- *Preventing document histories from leaking out through your firewall.*

Checking for document histories

Practical demonstration of checking OpenOffice.org documents for change information.

Further information

- A commented command history from the practical demonstrations will be published on <http://edb.jacob-sparre.dk/foredrag/00o/> after the talk.
- Write me at sparre@crs4.it if you have questions related to the talk.

The End.