Háttéradatok kigyűjtése extract és libextractor segítségével

Ne csak találgassuk a fájlok tulajdonságait kereséskor. Használjunk célzott kibontó bővítményeket a pontos fájladatbázis előállítására.

Modern állományformátumok leíró adatokban tárolják a fájlra vonatkozó információkat. Ezeket a fejlesztéseket az az igény hajtotta, hogy a fájlokat egyszerű állományneveken felül összetettebb módon is lehessen szervezni. Az ilyen háttéradatokkal az a probléma, hogy nem szabványos módon tárolódnak a különféle formátumokban. Ez sajnos megnehezíti az erősen formátumfüggő programok dolgát (ilyenek például a fájlkezelők vagy fájlmegosztó alkalmazások) ha hozzá szeretnének férni ehhez az információhoz. Következésképpen az adatok kibontására rengeteg elemző program látott napvilágot, ilyen jellegű többek közt az ***AVInfo***, az ***id3edit***, a ***jpeginfo*** és a ***Vocoditor***.

Cikkünkben a ***libextractor*** könyvtárat és kibontó eszközét mutatjuk be. A ***libextractor Project*** célja egységes felületet biztosítani különféle fájlformátumok háttéradatainak kisbontásához. A ***libextractor*** programot jelenleg az ***evidence***, a következő ***Enlightenment*** verzió fájlkezelője, valamint a ***GNUnet***, ez az név nélküli ellenőrizhetetlen peer-to-peer fájlmegosztó rendszer használja. Az ***extract*** eszköz a könyvtárhoz tartozó parancssoros kezelőfelület. A ***libextractor*** a ***GNU General Public License*** védelme alá tartozik.

A ***libextractor*** sok hasonlóságot mutat a népszerű ***file*** eszközzel, amely a fájl első bájtját használja a ***MIME*** típus kiderítéséhez. A ***libextractor*** ugyanakkor különbözik is a ***file***-tól, hiszen az egyszerű ***MIME*** típuson kívül sok egyéb információt is megpróbál kigyűjteni. A ***libextractor*** által kikeresett egyéb adatok közt találjuk az állomány létrehozására használt program nevét, az állomány szerzőjét, leírását, az albumcímeket, képméretet vagy éppen a film hosszát.

A ***libextractor*** ezeket az információkat a népszerű formátumokhoz írt értelmezők segítségével szerzi meg. A listán jelenleg ott találjuk az ***MP3***, ***Ogg***, ***Real Media***, ***MPEG***, ***RIFF (avi)***, ***GIF***, ***JPEG***, ***PNG***, ***TIFF***, ***HTML***, ***PDF***, ***PostScript***, ***Zip***, ***OpenOffice.org***, ***StarOffice***, ***Microsoft Office***, ***tar***, ***DVI***, ***man***, ***Deb***, ***ELF***, ***RPM***, ***asf*** formátumokat valamint az olyan általános megoldásokat mint a ***MIME***-típus érzékelése. Persze sok más formátum is létezik, de a népszerűbb formátumok közül csak néhány üzleti formátum nem támogatott.

Az új formátumok támogatása könnyen beépíthető, hiszen a ***libextractor*** bővítmények segítségével gyűjti össze az adatokat. A ***libextractor*** bővítmények tulajdonképpen megosztott könyvtárak amelyek általában egy adott formátum értelmezését végzik. Cikkünk végén azt is bemutatjuk, miképpen tudjuk egy új formátum támogatását könyvtárba szervezni.

A ***libextractor*** összegyűjti a különféle bővítményektől kapott háttéradatokat és az ügyfeleknek osztályozást és karaktersorozatot tartalmazó adat-pár listát ad vissza. Az osztályozási adatot a háttéradatok kategóriákba (cím, szerző, téma, leírás) szervezésére használjuk.

A libextractor telepítése és használata

A ***libextractor***-t legegyszerűbben úgy telepíthetjük, ha a terjesztésekben található bináris csomagot használjuk fel. ***Debian*** alatt az ***extract*** eszköz külön csomag, ***extract*** néven. Ha más alkalmazást szeretnénk ***libextractor***-hoz fordítani a ***libextractor0-devel*** csomagban találjuk a szükséges fejléceket. Amennyiben forrásból szeretnénk lefordítani a ***libextractort***, szokatlanul magas memóriaterületre lesz szükségünk: 256 MB rendszermemória nagyjából a minimum, ugyanis a ***GCC*** körülbelül 200 MB-ot használ az egyik bővítmény fordításánál. Egyébiránt a fordítás a megszokott utat követi, amint azt az *1. listában* láthatjuk.

A ***libextractor*** telepítése után, az ***extract*** eszköz segítségével gyűjthetjük ki a dokumentumok háttéradatait. Alapértelmezés szerint az ***extract*** eszköz adott bővítménykészletet használ, amely a ***libextractor*** jelenlegi verziójának valamennyi fájlformátum-jellegű bővítményét tartalmazza, a ***MIME***-típus érzékelő bővítménnyel együtt. A ***Linux Journal*** honlapról kapott kimenet a *2. listában* olvasható.

A ***BibTeX*** felhasználók számára valószínűleg jól jön a `-b` kapcsoló, amely a dokumentumokból automatikusan háttéradatokkal kiegészített ***BibTeX*** bejegyzéseket készít, amint azt a *3. listában* láthatjuk.

Másik érdekes kapcsoló a `-B LANG`. A kapcsoló az egyik nyelvfüggő, ám formátumfüggetlen bővítményt tölti be. Ezek a bővítmények megpróbálnak egyszerű szöveget keresni a dokumentumban kartersorozatokat összehasonlítva egy szótárállománnyal. Amennyiben furcsának tűnt a 200 MB-os memóriaigény a ***libextractor*** fordításakor, a választ ezek a bővítmények meg. a gyors szótárkeresés érdekében egy ***bloomfilter***-t hoz létre amely gyors valószínűség alapú egyezés keresést tesz lehetővé; a létrejövő adatstruktúrát a *GCC* némiképp nehezen nyeli le.

1. lista A libextractor fordítása 200MB memóriát igényel

```
$ wget
http://ovmj.org/libextractor/download/libextrac
tor-0.4.1.tar.gz
$ tar xvfz libextractor-0.4.1.tar.gz
$ cd libextractor-0.4.1
$ ./configure –prefix=/usr/local
$ make
# make install
```

2. lista Háttéradatgyűjtés HTML-ből

```
$ wget -q http://www.linuxjournal.com/
$ extract index.html
description - The Monthly Magazine of the Linux
Community
keywords - linux, linux journal, magazine
```

3. lista BibTeX bejegyzések készítése igen egyszerű, ha a dokumentum számos hátéradattal rendelkezik

```
$ wget -q
http://www.copyright.gov/legislation/dmca.pdf
$ extract -b ~/dmca.pdf
% BiBTeX file
@misc{ unite2001the_d,
    title = "The Digital Millennium Copyright
    ➥Act
    of 1998",
    author = "United States Copyright Office
    ➥- jmf",
    note = "digital millennium copyright act
    circumvention technological protection
    ➥management
    information online service provider
    ➥liability
    limitation computer maintenance competition
    repair ephemeral recording webcasting
    ➥distance
    education study vessel hull",
    year = "2001",
    month = "10",
    key = "Copyright Office Summary of the
    ➥DMCA",
    pages = "18"http://www.netpincer.hu/
    ➥index.php
}
```

4. lista A libextractor néha még ismeretlen formátum esetén is használható információkat nyerhet ki

```
$ wget -q http://www.bayern.de/HDBG/polges.doc
$ extract -B de polges.doc | head -n 4
unknown - FEE Politische Geschichte Bayerns
Herausgegeben vom Haus der Geschichte als Heft
der zur Geschichte und Kultur Redaktion Manfred
Bearbeitung Otto Copyright Haus der Geschichte
München Gestaltung fürs Internet Rudolf Inhalt
im.
unknown - und das Deutsche Reich.
unknown - und seine.
unknown - Henker im Zeitalter von Reformation
          ➥und Gegenreformation.
```

A `-B` kapcsoló még nem dokumentált vagy támogatott formátumok esetén jöhet jól. A ***printable*** bővítmények általában a dokumentum teljes szövegét sorban kinyomtatják. A ***4. lista Microsoft Word*** dokumentumra alkalmazott ***extract*** futási eredményét mutatja be.
A németül értők láthatják, hogy a kapott anyag elég jó leírása a szövegnek. A támogatott nyelv még a ***Dán (da)***, ***Német (de)***, ***Angol (en)***, ***Spanyol (es)***, ***Olasz (it)*** és a ***Norvég (no)***. Más nyelvek támogatása mindössze a megfelelő karakterkészlettel beillesztett ingyenes szótárkészleten múlik.
A többi kapcsolót az ***extract*** kézikönyvoldala ismerteti; lásd a `man 1 extract` parancsot.

libextractor saját projektjeinkben

Az *5. lista* egy ***libextractor*** könyvtárat használó minimális programot mutat be. A ***minimal.c*** fordításához a `-lextractor` kapcsolót kell megadnunk a ***GCC***-nek. Az `EXTRACTOR_KeywordList` egyszerű láncolt lista, amelyben a kulcsszavak és a kulcsszó típusokat találjuk. A részleteket, bővítmények betölt és a kulcsszólisták kezelését végző további függvények leírását a ***libextractor*** kézikönyvoldalán találjuk: `man 3 libextractor`. A ***Java*** programozóknak érdemes tudni, hogy a ***libextractorral JNI*** segítségével kommunikálni képes ***Java*** osztály is létezik.

Bővítmények készítése

A ***libextractor*** bővítmények készítésekor a legnehezebb feladat az adott formátumhoz tartozó értelmező megírása. Mindazonáltal az eljárás lényegében szinte mindig ugyanaz. A bővítmény könyvtárat ***libextractor_XXX.so*** néven kell elkészíteni, ahol az ***XXX*** a bővítmény által kezelt fájlformátumot jelzi. A könyvtárból exportálni kell a ***libextractor_XXX_extract*** metódust a *6. listában* bemutatott definíció szerint.
A `filename` paraméter a feldolgozandó állomány nevét adja meg. A `data` a fájl általában ***mmap***-olt formában tárolt adataira mutat, végül a size adja meg az állomány méretét. A legtöbb bővítmény (plugin) nem használja fel az állomány nevét és közvetlenül az adatokat elemezi, rögtön az első lépésben ellenőrizve, hogy az állomány fejléce egyezike az adott formátum által igényelt alakkal.

5. lista minimal.c egybegyűjtve mutatja be a legfontosabb libextractor függvényeket

```
#include <extractor.h>
int main(int argc, char * argv[]) {
  EXTRACTOR_ExtractorList * plugins;
  EXTRACTOR_KeywordList   * md_list;
  plugins = EXTRACTOR_loadDefaultLibraries();
  md_list = EXTRACTOR_getKeywords(plugins,
  ➥argv[1]);
  EXTRACTOR_printKeywords(stdout, md_list);
  EXTRACTOR_freeKeywords(md_list);
  EXTRACTOR_removeAll(plugins); /* unload
  ➥plugins */
}
```

A `prev` az eddigi bővítmények által kigyűjtött kulcsszavak listája. A függvénynek a kulcsszólista frissített változatát kell visszaadnia. Amennyiben a formátum nem esik egybe a bővítmény által várttal, a `prev` adódik vissza. A legtöbb bővítmény az `addKeyword` vagy hasonló függvényekkel bővíti a listát (*7. lista*).
Az `addKeyword` leggyakrabban a ***MIME*** típus felvételére használják miután a fájl fajtája már kiderült. Például a ***JPEG-extractor*** (*8.lista*) ellenőrzi a ***JPEG*** fejléc első bájtjait és vagy megszakítja futását vagy úgy dönt az állomány valóban ***JPEG*** típusú. A kódban található `strdup` fontos hiszen a karaktersorozat memóriafoglalása később megszűnhet (általában az `EXTRACTOR_freeKeywords()` hívás során). A támogatott kulcsszó osztályokat, azaz példánkban az `EXTRACTOR_MIMETYPE` típust, az `extractor.h` fejlécállományban találjuk.

Összefoglalás

A ***libextractor*** egyszerű bővíthető ***C*** könyvtár, amely képes összegyűjteni a dokumentumok háttéradatait. Bővítményalapú szerkezete és széleskörű formátumismerete kiemeli a formátumfüggő eszközök közül. A ***libextractor*** ugyanakkor korlátozott a tekintetben, hogy a nem képes módosítani a háttéradatokat, amelyet egyébként a specializálódott eszközök általában képesek megtenni.

Linux Journal 2005. június, 134. szám

6. lista A libextractor bővítményből exportálandó függvény definíciója

```
struct EXTRACTOR_Keywords *
libextractor_XXX_extract
   (char * filename,
    char * data,
    size_t size,
    struct EXTRACTOR_Keywords * prev);
```

7. lista A bővítmények egyszerű láncolt listaként adják vissza a háttéradatokat

```
static void addKeyword
   (struct EXTRACTOR_Keywords ** list,
    char * keyword,
    EXTRACTOR_KeywordType type)
{
  EXTRACTOR_KeywordList * next;
  next = malloc(sizeof(EXTRACTOR_KeywordList));
  next->next = *list;
  next->keyword = keyword;
  next->keywordType = type;
  *list = next;
}
```

8. lista A fájl fejlécének értelmezése után a jpegextractor.c felveszi a MIME típust a listába

```
if ( (data[0] != 0xFF) || (data[1] != 0xD8) )
  return prev; /* not a JPEG */
addKeyword(&prev,
           strdup("image/jpeg"),
           EXTRACTOR_MIMETYPE);
/* ... további értelmező kód ... */
return prev;
```

KAPCSOLÓDÓ CÍMEK

- gnunet.org/libextractor
- getid3.sf.net
- evidence.sf.net
- ovmj.org/GNUnet
- www.wotsit.org
- dublincore.org/documents/dcmi-terms
- dmoz.org/Computers/Software/Typesetting/TeX/BibTeX
- enlightenment.org
- ovmj.org/GNUnet/download/bloomfilter.ps

Christian Grothoff
2000-ben diplomázott a Wuppertali Egyetemen matematika szakon. Jelenleg PhD hallgató a Purdue Egyetemen, ahol a statikus programanalízis és biztonságos peer-to-peer hálózatkezelés témakörét tanulmányozza. 1995 óta Linux felhasználó, több nyílt program fejlesztésében is segédkezett, jelenleg a GNUnet karbantartója és a libextractor magcsapatának tagja. Honlapja a grothoff.org/christian címen érhető el.