PROCESSING ON THE CHM FORMAT

Pavel KONEČNÝ, Bachelor Degree Programme (3)
Dept. of Intelligent Systems, FIT, BUT
E-mail: xkonec36@stud.fit.vutbr.cz

Supervised by: Dr. Petr Hanáček

ABSTRACT

The aim of this work is the analysis of CHM (Compiled HTML), used initially as a help file in Microsoft Windows operating systems and the implementation of unwrapping files that were contained in this format. Creating this file is the opposite if its extraction, so-called compiling. The aim is to focus on unwrapping the encapsulated data, which can be compressed using LZX compression and covering the remaining part of the format. This is useful for common usage of the CHM file.

The implementation of unwrapping is created in C++ and the code was designed so that it is not dependant on the platform.

1 ÚVOD

CHM (Compiled HTML) formát souboru, který je také někdy nazývaný „HTML Help“, byl vyvinut společností Microsoft jako soubor nápovědy pro používání v rodině operačních systémů rodiny Microsoft Windows.

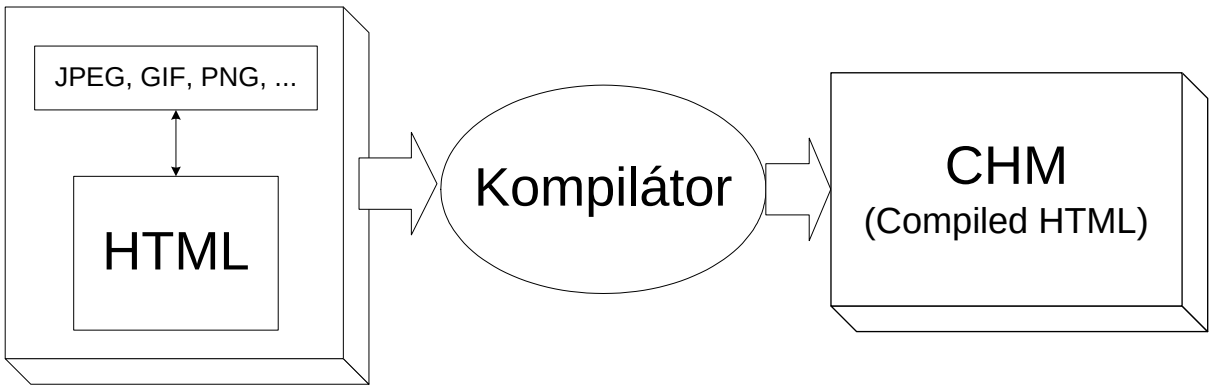
Tato práce se zabývá opačným procesem než je tvorba tohoto souboru. Hlavním cílem je získat data (soubory), které byly předlohou pro jeho vytvoření. Tento problém komplikuje skutečnost, že není dostupný žádný oficiální, úplný a přesný popis formátu.

2 SPECIFIKACE FORMÁTU

CHM soubor v sobě zapouzdřuje data různých formátů a zachovává přitom adresářovou strukturu dat. To umožňuje zakomponovat do takového souboru HTML stránky, které mohou mít odkazy na jiné soubory, a ty jsou právě často zabaleny ve stejném CHM souboru. Celá struktura HTML stránek může být proto obsažena v jediném souboru a v tom spočívá hlavní výhoda tohoto formátu.

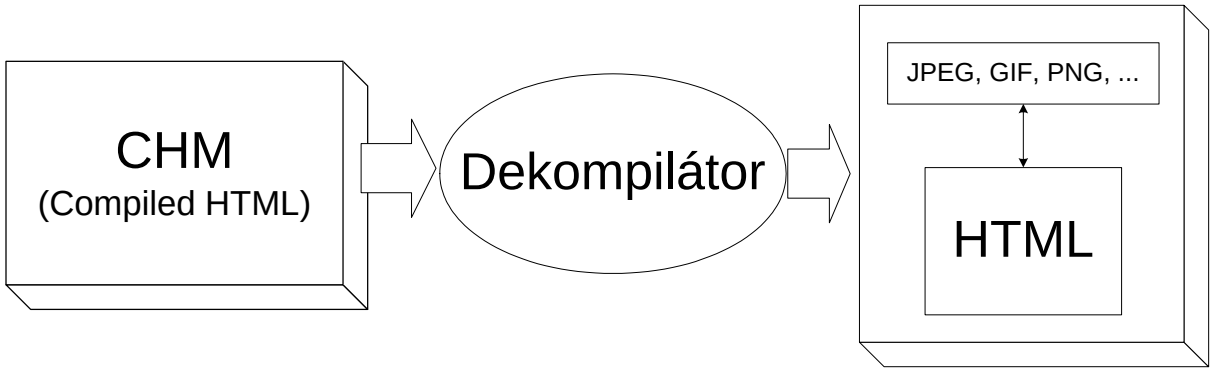
Soubor je vytvořen pomocí tzv. kompilátoru (jak napovídá zkratka souboru CHM – „Compiled HTML“). Obrázek 1 nám zjednodušeně znázorňuje vznik CHM souboru. Ve skutečnosti jsou ve výsledku obsažena i data, která mohou například definovat výrazy pro indexové vyhledávání, obsah obsažených souborů uložený v předepsané stromové struktuře, aj.

Formát CHM umožňuje na zapouzdřená data použít kompresi LZX (Lempel-Ziv eXtended). Jedná se o neztrátovou kompresi, kterou používá společnost Microsoft pro vytváření tzv. „kabinet“ souborů (CAB). Obdoba těchto archívů je použita i v CHM formátu podle předpisu udávajícího, kde budou uloženy informace o kompresi a kde samotná komprimovaná data.



Obr. 1: *Zjednodušené schéma tvorby CHM souboru*

„Dekompilací“ by se dal nazvat opačný proces než je vytváření souboru CHM. Samotné implementaci získávání dat předchází analýza formátu. Přitom jsou vyzdviženy důležité části pro získání dat a jsou zastíněny ty, které nejsou svým obsahem důležité pro tuto problematiku. Hlavní podstata tzv. „dekompilace“ je znázorněna na Obrázku 2.



Obr. 2: *Zjednodušené schéma “dekompilace” CHM souboru*

Pro získání požadovaného obsahu, kterým jsou data, která byla vstupem při vytváření souboru CHM, je první důležitou fází získání seznamu souborů – tzv. adresáře. Ten je hlavním podkladem pro další získávání obsahu, který je uložen v sekcích. Sekce mohou být zkomprimovány a v tom případě je třeba nejdříve získat podrobnosti o kompresi a rozbalit tak příslušnou sekci pro získání dat.

3 IMPLEMENTACE

Součástí této práce a taky hlavním programovým cílem je knihovna, která zprostředkovává zpřístupnění dat z CHM formátu, kde je hlavním cílem, aby nezáleželo na verzi právě zpracovávaného souboru. Získávání dat touto knihovnou probíhá v několika hlavních krocích:

- inicializace souboru, který má být zpracován,
- zpracování důležitých částí souboru, jehož hlavním výstupem je seznam souborů (tzv. adresář souborů), pomocí kterého lze pak získávat příslušná data,
- získávání určených souborů (dat) pro jejich další zpracování.

Poslední bod pak může být rozdělen na několik fází, které závisí na způsobu uložení dat. Ty jsou však zastíněny z pohledu užívání knihovny a probíhají automaticky bez zásahu uživatele. Jsou to tyto fáze:

- na základě adresáře souborů se zjistí sekce, v jaké jsou data umístěna, pozice v této sekci a velikost těchto požadovaných dat,
- pokud je sekce komprimována, je nutné nejdříve zjistit informace o použité kompresi, které jsou důležité pro dekompresi sekce. Teprve poté je možné rozbalit sekci, nebo její části, kde jsou data uložena,
- přečtení získaných dat, která byla požadována.

4 ZÁVĚR

Implementace knihovny je napsána v přenositelném kódu, který splňuje všechny požadavky na dekompozici CHM souboru. Uplatnění může nalézt v mnoha případech, jako jsou i například kontroly souborů antivirovými programy nebo rekonstrukce dat, která byla původně podkladem pro vytvoření CHM.

Jedním z hlavních faktorů, na které byl brán zřetel při tvorbě, byla efektivita, kterou může ovlivnit jak rychlost zpracování, tak například velikost využívané paměti apod. Pro získání co nejlepších vlastností bylo zapotřebí otestovat program na různých vzorcích souborů, ať už na těch, které jsou běžně používány, nebo na speciálně vytvořených souborech za účelem zátěžového testování.

LITERATURA

[1] Russotto, M. T.: HTML Help (CHM) Tools and Information, 2003, Dokument dostupný na URL `http://www.speakeasy.org/~russotto/chm/` (březen 2005)

[2] Microsoft Corporation: Microsoft LZX Data Compression Format, 1997, Dokument dostupný na URL `http://download.microsoft.com/download/platformsdk/cab/2.0/w98nt42kmexp/en-us/cabsdk.exe` (březen 2005)