Rövid betekintés a GEDCOM állományok felépítésébe

Dr. Hatvany Béla Csaba, Kuchen, Németország (csaba@hatvany.de)

A zenésznek illik egy keveset olaszul tudni – szokták mondani. Az informatikusnak illik egy keveset angolul tudni – ezt ma mindenki elfogadja. A genealógusnak illik egy keveset GEDCOM-ul tudni – állítom én. Nézzük meg közelebbről, miről is van szó?

A GEDCOM mozaikszót a GEnealogical Data COMmunication (genealógiai adatközvetítés) angol kifejezésből vett betűkből rakták össze. A mozaikszóval egy állományformátumot, illetve egy szabványt szoktunk megnevezni. A GEDCOM formátumú állományok kiterjesztése `.ged`. A GEDCOM szabványt az Utolsó Napok Szentjeinek Jézus Krisztus Egyházán (The Church of Jesus Christ of Latter-day Saints, "Mormon Egyház") belül működő Családtörténeti Osztály (Family History Department) informatikusai hozták létre, azzal a céllal, hogy elősegítsék és megkönnyítsék a genealógiai adatok rögzítését, közvetítését és összehasonlítását.

A specifikáció első verziója 1984-ben látott napvilágot. Azóta sokat változott és a jelenleg érvényes 5.5 verziót, amit 1995-ben fogadtak el, manapság minden valamire való családfaprogram támogat. Komoly családkutatást ma már nem is végezhetünk a GEDCOM szabvány nélkül: a hatalmas adatmennyiség, amely ma rendelkezésünkre áll, nem kezelhető egy megfelelően szabványosított adatstruktúra nélkül. Napjainkban ezt a szerepet a GEDCOM szabványt tölti be. Valamikor a távoli jövőben, amikor a világ összes anyakönyve, temetője, levéltári állománya, stb. digitalizálva lesz, lehet, hogy a kutatók áttérnek majd egy másik szabvány használatára. Addig viszont érdemes megismerkedni a GEDCOM állományformátum alapjaival.

- A GEDCOM ismerete nélkül nem tudom a családom történetét kutatni? – hallom a kétkedő kérdést. Dehogynem, de ismerve a GEDCOM formátumot könnyebben és hatékonyabban kezelhetjük adatainkat és – nem túlzás – új adatokra is hamarabb szert tehetünk. Ezt kívánom megvilágítani ebben és a következő írásomban.

Egy GEDCOM állomány egy egyszerű szövegállomány, amit egy közönséges szövegszerkesztő programmal is megtekinthetünk. Az állományban hosszabb-rövidebb szövegsorokat látunk. Ezeknek az a közös tulajdonságuk, hogy három oszlopot különböztethetünk meg: az első oszlop (amelyiket piros színnel emeltem ki) egész számokat, a második oszlop (amelyiket kék színnel emeltem ki) négy betűs címkéket, míg a harmadik oszlop (amelyiket zöld színnel emeltem ki) hosszabb vagy rövidebb szöveges információt tartalmaz.

```
0 HEAD          ←
1 SOUR GenoPro ← a GEDCOM állományt létrehozó program
2 NAME GenoPro®          ← a program leírása
2 VERS 2.5.3.8 ← a program verziója
2 CORP GenoPro ← a programot gyártó cég
2 ADDR http://www.genopro.com ← a cég honlapja
1 DATE 9 JUL 2011       ← az állomány létrehozásának dátuma
1 CHAR UTF-8       ← a használt karakterkészlet
1 GEDC          ← az állomány formátuma
2 VERS 5.5          ← az alkalmazott GEDCOM verzió
2 FORM LINEAGE-LINKED          ← az adatstruktúra típusa
```

A negyedik oszlop (amelyiket *dőlt szedéssel* írtam és egy ←jellel indítottam a sorokat) nem tartozik a GEDCOM állományhoz: ez elmagyarázza a sorokban lévő információt és én adtam hozzá az állományhoz.

Az első oszlopban levő számok az információszintet adják meg. A legmagasabb szintet a 0 szám jelzi, a többi szintek alárendelt szintek. A 0 szinttel egy *adatsor* vagy *adatrekord*, vagy egyszerűen egy *rekord* kezdődik és addig tart, ameddig egy 0-val bevezetett újabb rekord nem kezdődik el. A rekordot bevezető 0 után egy rekordazonosító és a rekordtípus áll. A fenti példában csak egy rekordot látunk, melynek típusa HEAD és nincs rekordazonosítója.

A GEDCOM szabvány többféle rekordot ismer. Mi csak a legfontosabb rekordtípusokat fogjuk megismerni. A fenti példa a HEAD típusú rekordot mutatja be, ami a GEDCOM állomány fejléce. A fejléc fontos információt tartalmaz az állományról. Erre az információra szüksége van a GEDCOM állományt feldolgozó programnak. A fenti példában ez az információ háromszintű. Minden szint a közvetlenül fölötte elhelyezkedő magasabb szintre vonatkozik. Bonyolultnak hangzik, de valójában nagyon egyszerű. Lássuk csak közelebbről!

A második oszlop négytagú címkéi (angolul *tag*) leírják az általuk megadott információt. Így az első szintű SOUR címke megadja a közvetlen felette levő magasabb szintű információ (azaz a 0 szintű HEAD. más szóval a szóban forgó állomány) forrását (*source*). A mi állományunkat tehát a GenoPro program hozta létre. A következő második szintű NAME címke megadja a közvetlen felette levő magasabb szintű információ (azaz az állományunkat létrehozó program) nevét vagy rövid leírását: GenoPro®. Az ezután következő szintén második szintű VERS címke szintén a programra vonatkozik és megadja a verzióját (*version*): 2.5.3.8. Ugyanígy a CORP és ADDR címkék által megadott információ szintén a programra vonatkozik: cég- és internetcímet tudunk meg. Az ezután következő három első szintű információ viszont nem a programra (mely szintén első szintű), hanem magára az

állományra vonatkozik. Sorra megtudjuk, mikor hozták létre az állományt (DATE), hogy milyen karakterkészletet használtak (CHAR) és hogy egy GEDCOM formátumú állományról van szó (GEDC). A GEDCOM formátumot a szabvány 5.5 verziója szerint alkalmazták (VERS) és az adatstruktúra típusa eredetkapcsolat (FORM) (*lineage-linked*, mindegyik GEDCOM állományban az adatstruktúra ilyen típusú; ennek részletezése meghaladja ennek az írásnak a kereteit).

Amint látjuk, a HEAD típusú rekordot inkább a feldolgozó programnak, mintsem a családkutatónak nyújt információt, noha sokszor a kutató is hasznát veheti ezeknek az adatoknak. Sokkal érdekesebb az INDI típusú rekord felépítése, melyből alább láthatunk egy példát.

```
0 @ind00008@ INDI
1 NAME Mihály /Ásbóth/
2 GIVN Mihály
2 SURN Ásbóth
2 ALIA János
1 SOUR @source00003@
1 SEX M
1 BIRT
2 DATE 1790
2 PLAC Sopron
2 SOUR @source00003@
1 DEAT
2 DATE 25 Apr 1872
2 PLAC Eperjes
2 SOUR @source00001@
1 NOTE Asbóth Mihály Jánosnak nevezik mindenhol
2 CONT Egy évtizeden át geodétaként dolgozott a
2 CONC Dunán, Visegrád környékén és Felsővisó k
2 CONC örzetében, Bustyaházán; a Tisza felső sz
2 CONC akaszán hidakat, terelőműveket, utakat t
2 CONC ervezett és épített. Ő építette a márama
2 CONC rosszigeti sókikötőt (1844). Oklevelét a
2 CONC  pesti mérnöki intézetben 1816-ban kapta
2 CONC  1826-ig geodétaként dolgozott a Dunán,
2 CONC  Visegrád környékén, valamint Felsővisó
2 CONC körzetében. 1830 körül kamarai építésszé
2 CONC  nevezték ki.Kéziratos térképei és jelen
2 CONC tései az Orsz. Levéltárban vannak.
1 OCCU @occu00004@
1 FAMS @fam00007@
1 FAMC @fam00498@
```

Az INDI típusú rekordok egy egyén (*individual*) adatait rögzítik. A rekordban látunk két @ között elhelyezett kifejezést. Ezek azonosítók. A mi egyénünknek az állományban a @ind00008@ azonosító felel meg. Az azonosító kifejezések rendszerint egy rövidítésből (`ind`, `fam`, `souce`, stb.) és egy számból tevődnek össze.

Példánkban ez egyén neve (**NAME**) Ásboth Mihály. A név utónév/utónevek-vezetéknév sorrendben szerepel, a vezetéknév mindig két ferdevonal (/) között van. A rekord még egyszer rögzíti az utónevet (**GIVN**, *given name*) és a vezetéknevet (**SURN**, *surname*). Az **@ind00008@** azonosítóval rendelkező személy adatait a kutató a **@source00003@** azonosítóval rendelkező forrásban (**SOUR**, *source*) találta meg.

A kutatott személyünk férfi (**SEX**), született (**BIRT**, *birth*) 1790-ben (**DATE**) Sopronban (**PLAC**, *place*) és meghalt (**DEAT**, *death*) 1872. április 25-én (**DATE**) Eperjesen (**PLAC**). A születés adatait a kutató a **@source00003@** azonosítóval rendelkező forrásból, míg a halál adatait a **@source00001@** azonosítóval rendelkező forrásból (**SOUR**) merítette.

A **NOTE** címke alatt egy megjegyzést fűzhetünk a kutatott személyhez, melynek tartalmát a **CONT** (*content*) címke alatt adhatjuk meg és a **CONC** (*concatenation*) címke segítségével írhatjuk több soron át. A foglakozásra vonatkozó adatokat az **OCCU** (*occupation*) címke alatt adhatjuk meg.

Az utolsó két címke a családi állapotra vonatkozik. A **FAMS** (*family-spouse*) címke azonosítja a családot, amelyben a szóban forgó személy házastársként szerepel, míg a **FAMC** (*family-child*) címke azt a családot azonosítja, amelyikben a szóban forgó személy gyerekként szerepel. A két családrekord esetünkben ez:

```
0 @fam00007@ FAM
1 HUSB @ind00008@
1 WIFE @ind01714@
1 CHIL @ind00022@
1 CHIL @ind00167@
1 CHIL @ind00184@
```

illetve ez

```
0 @fam00498@ FAM
1 HUSB @ind01711@
1 WIFE @ind01712@
1 CHIL @ind01713@
1 CHIL @ind00008@
1 CHIL @ind00098@
```

Mindkét esetben egy egyszerű családrekordot látunk. Az első családrekord szerint a férj (**HUSB**, *husband*) az **@ind00008@** azonosítót viselő személy, azaz Ásboth Mihály, míg a feleség (**WIFE**) az **@ind01714@** azonosítót viseli. A rekordból kiderül még, hogy Ásboth Mihálynak három gyermeke (**CHILD**) volt. A **@fam00498@** azonosítóval rendelkező családrekordból, mely **FAMC** címkével van Ásboth Mihály egyéni rekordjához kötve, kiderül, hogy a geodétának két testvére volt.

Természetesen, itt nem térhetek ki a GEDCOM szabvány részletes és kimerítő ismertetésére. Az érdeklődő Olvasó itt: http://homepages.rootsweb.ancestry.com/~pmcbride/gedcom/55gctoc.htm megtalálja a szabvány hivatalos angol szövegét és szükség esetén ott kereshet eligazítást.

E rövid betekintés a GEDCOM szabvány struktúrájába csak felületén érintheti e terjedelmes témát. De a GEDCOM állományok felépítésének még ilyen felületes ismerete is segít a genealógiai adatok kezelésében és esetenként akár új adatokhoz is elvezethet. Hogy miként? – azt a következő írásomban mondom el.