



Digitalizált törvényhozási tudástár 2. (EKOP-2.1.24-2013-2013-0001)

A KÖZBESZERZÉS MŰSZAKI LEÍRÁSA

MŰSZAKI SPECIFIKÁCIÓ

1. A szerződéshez rendelt rövid elnevezés:

DTT2 projekt keretében digitalizálásra kerülő országgyűlési dokumentumokkal kapcsolatos feladatok.

2. A projekt célja:

Az országgyűlési dokumentumok digitalizált anyaga legyen alkalmas arra, hogy betöltésre kerülhessen a Parlamenti Információs Rendszerbe (PAIR) és az Országgyűlési Könyvtár katalógusába.

3. Általános elvek:

- 3.1 A digitalizálás során a dokumentumok hitelességének megőrzése és a visszakereshetőségük biztosítása. Ezt a kétrétegű PDF technika biztosítja a legcélszerűbben és legköltséghatékonyabban.
- 3.2 A dokumentumok (irományok és bizottsági jegyzőkönyvek) jelentős része a dokumentum fizikai egységének megőrzése nélkül digitalizálható (a dokumentumok lapra szedhetők, laponként digitalizálhatóak), kivéve egyes dokumentumokat, ezekről Ajánlatkérő külön listát készít. Ennek nagyságrendje nem éri el az egész mennyiség 10 %-át.

4. A közbeszerzés tárgya és mennyisége:

Az országgyűlési dokumentumok két nagy típusának, irományoknak és bizottsági jegyzőkönyveknek a digitalizálása, karakterfelismertetése (OCR), kétrétegű PDF készítése, a specifikációban megadott metaadatok számítógépes rögzítése.

Teljes mennyiség:

- Irományok 1992-2003 közötti időszak, 567.000 + 10% oldal terjedelemben.
- Bizottsági jegyzőkönyvek 1990-2006 közötti időszak, 1.070.000 + 10% oldal terjedelemben.

5. Módszertan – kétrétegű PDF készítés:

A kétrétegű PDF tulajdonságai:

- *Az első réteg:* a dokumentum írásképet tartalmazza, úgy ahogyan a dokumentum a digitalizálás időpontjában látható. Ez gyakorlatilag a dokumentumról készült fénykép.
- *A második réteg:* az optikai karakterfelismerésen átesett szöveg. Ez a szöveg az alapja a keresésnek. Ajánlatkérő meghatározza a karakterfelismerés (OCR) általános minőségi követelményét (minimum 95 százalékos pontosság) és a szövegfeldolgozás speciális minőségi követelményét (minimum 999 ezrelékes pontosság).

6. A közbeszerzés keretében megvalósuló tevékenységek:

6.1. Irományok:

Az irományok szkennelése, karakterfelismertetése (OCR), kétrétegű PDF/A-1 (ISO 19005-1:2005) formátum készítése, a specifikációban megadott metaadatok számítógépes rögzítése 567.000 oldal terjedelemben.

6.1.1. Szkennelési követelmények:

- A kívánt felbontás 300 DPI.
- A kívánt képformátum multipage TIFFCCITT Group 4 tömörítéssel.
- Szöveges dokumentumok esetében a színmélység 1 bit.
- A színes dokumentumok (részek) esetében a színmélység 24 bit (multipage TIFF LZW/ZIP tömörítéssel). Ajánlatkérő a dokumentumok Kísérőlapjain külön jelzi a színesben kért oldalak helyét.

6.1.2. Az oldal képminősége:

- Az egy-egy dokumentumoldarról készített digitális felvételnek egy képnek kell lennie, nem lehet sem szerkesztett, sem összeillesztett kép.
- Egy kép csak egy dokumentumoldalt tartalmazhat.
- A képnek a teljes oldalt kell tartalmaznia.
- A nyomtatott oldal mérethűen, az eredeti dokumentumnak megfelelően (1:1 arányban) legyen elmentve.
- Az oldalméret beállítása dokumentumonként legyen azonos (az oldalméret közötti eltérés mértéke mind szélességben, mind hosszúságban ne haladja meg a +/- 5 pixelt. Az egyes dokumentumokban előforduló eltérő méretű dokumentumrészekről a Kísérőlap ad iránymutatást.
- A dokumentumok oldaltájolását megváltoztatni nem szabad.
- Ajánlatkérő a következő képjavító eljárás elvégzését kéri: körbevágás.

6.1.3. A szkennelés módja:

- A dokumentum fizikai egységének megőrzése nélkül (lapra szerelve) kerül sor a digitalizálásra. Ajánlatkérő a dokumentumok Kísérőlapjain külön jelzi, ha a dokumentum fizikai egységét meg kell őrizni a szkennelés során.
- Ajánlatkérő elvárja, hogy a Vállalkozó a digitalizálási eljárás során a legnagyobb gondossággal járjon el. Ha az Ajánlattevő a dokumentumon olyan súlyos sérülést észlel, amely érdemben befolyásolja a műszaki paraméterek teljesülését, vagy a munkafolyamat a dokumentum további súlyos károsodását okozná, akkor köteles egyeztetést kezdeményezni az Ajánlatkérővel.

6.1.4. A szkennelés ütemezése:

Ajánlatkérő a digitalizálás ütemezésére a Vállalkozótól ütemtervet kér.

6.1.5. A számítógépesen rögzítendő metaadatok köre:

- Iromány száma
 - Iromány címe
 - Benyújtás dátuma
 - Benyújtó/k országgyűlési adatbázisbeli kódja
 - Digitalizált fájl neve az országgyűlési adatbázis számára
 - Digitalizált file neve a könyvtár számára
- A fenti adatokat a dokumentumok tartalmazzák, azokról rögzíthetők, illetve a benyújtók kódját az Ajánlatkérő a Vállalkozónak átadja (név, kód formában, amennyiben szükséges, akkor egyéb nyilvános adatokat is).
 - A szövegfeldolgozás pontosságának ebben a körben minimum 999 ezrelékesnek kell lennie (azaz 1000 karakterből 1 lehet hibás).

6.1.6. Adatbázisba szervezés:

- Ajánlatkérő a digitalizált állományokat adatbázisba szervezve kéri, azaz a digitális adatbázisban kereshetően álljanak rendelkezésre a metarekordok.
- Az adatbázisba szervezéshez szükséges egy CSV (Comma-separated values) fájl, amely tartalmazza a következő adatokat.

| Iromány száma | Iromány címe | Benyújtás dátuma | Benyújtó/k országgyűlési adatbázisbeli kódja | Digitalizált fájl neve |
|---------------|--------------|------------------|--|------------------------|
| | | | | |

A szöveges mezők adattartalmát idézőjelek ("doublequote") közé kell tenni.

Több benyújtó esetén a neveket a mezőn belül függőleges vonallal: |(UTF-8 name: PIPE, code: U+01C0) kell elválasztani.

6.2. Bizottsági jegyzőkönyvek:

A bizottsági jegyzőkönyvek szkennelése, karakterfelismertetése (OCR) és kétrétegű PDF/A-1 (ISO 19005-1:2005) formátum készítése, az MS-Word formátumú fájlok PDF/A-1 (ISO 19005-1:2005) formátumba konvertálása, a specifikációban megadott metaadatok számítógépes rögzítése 1.070.000 +/- 5 % oldal terjedelemben.

6.2.1. Szkennelési követelmények:

- A kívánt felbontás 300 DPI.
- A kívánt képformátum multipageTIFFCCITT Group 4 tömörítéssel.
- Szöveges dokumentumok esetében a színmélység 1 bit.
- A színes dokumentumok (részek) esetében a színmélység 24 bit (multipage TIFF LZW/ZIP tömörítéssel). Ajánlatkérő a dokumentumok Kísérőlapjain külön jelzi a színesben kért oldalak helyét.

6.2.2. Az oldal képminősége:

- Az egy-egy dokumentumoldalról készített digitális felvételnek egy képnek kell lennie, nem lehet sem szerkesztett, sem összeillesztett kép.
- Egy kép csak egy dokumentumoldalt tartalmazhat.
- A képnek a teljes oldalt kell tartalmaznia.
- A nyomtatott oldal mérethűen, az eredeti dokumentumnak megfelelően (1:1 arányban) legyen elmentve.
- Az oldalméret beállítása dokumentumonként legyen azonos (az oldalméret közötti eltérés mértéke mind szélességben, mind hosszúságban ne haladja meg a +/- 5 pixelt. Az egyes dokumentumokban előforduló eltérő méretű dokumentumrészekről a Kísérőlap ad iránymutatást.
- A dokumentumok oldaltájolását megváltoztatni nem szabad.
- Ajánlatkérő a következő képjavító eljárás elvégzését kéri: körbevágás.

6.2.3. A szkennelés módja:

- A dokumentum fizikai egységének megőrzése nélkül (lapra szerelve) kerül sor a digitalizálásra. Ajánlatkérő a dokumentumok Kísérőlapjain külön jelzi, ha a dokumentum fizikai egységét meg kell őrizni a szkennelés során.
- Ajánlatkérő elvárja, hogy a Vállalkozó a digitalizálási eljárás során a legnagyobb gondossággal járjon el. Ha az Ajánlattevő a dokumentumon olyan súlyos sérülést észlel, amely érdemben befolyásolja a műszaki paraméterek teljesülését, vagy a munkafolyamat a dokumentum további súlyos károsodását okozná, akkor köteles egyeztetést kezdeményezni az Ajánlatkérővel.

6.2.4. A szkennelés ütemezése:

- Ajánlatkérő a digitalizálás ütemezésére a Vállalkozótól ütemtervet kér.

6.2.5. MS-Word dokumentumok konvertálása:

- Az Ajánlatkérő átadja a bizottsági jegyzőkönyvek azon körét, amely MS-Word formátumban ("Word 5.0 for DOS" verziótól "MS-Word 97" verzióig) rendelkezésre áll.
- Az átadott fájlokat az Ajánlattevő feladata konvertálni PDF/A-1 (ISO 19005-1:2005) formátumba.

6.2.6. A számítógépesen rögzítendő metaadatok köre:

- Bizottság országgyűlési adatbázisbeli kódja
- Ülés dátuma
- Ülés száma
- Digitalizált fájl neve

- A fenti adatokat a dokumentumok tartalmazzák, azokról rögzíthetőek, illetve a bizottságok kódját az Ajánlatkérő a Vállalkozónak átadja (ciklus, név, kód formában, amennyiben szükséges, akkor egyéb nyilvános adatokat is).
- A szövegfeldolgozás pontosságának ebben a körben minimum 999 ezrelékesnek kell lennie, (azaz 1000 karakterből 1 lehet hibás.).

6.2.7. Adatbázisba szervezés:

Általános követelmények:

- Ajánlatkérő a digitalizált állományokat xml fájlba szervezve kéri, azaz azokban kereshetően álljanak rendelkezésre a metarekordok.
- A fájlba szervezéshez szükséges egy CSV (Comma-separated values) fájl, amely tartalmazza a következő adatokat.

| | | | |
|--|----------------|---------------|---------------------------|
| Bizottság országgyűlési adatbázisbeli kódja | Ülés dátuma | Ülés száma | Digitalizált fájl neve |
|--|----------------|---------------|---------------------------|

A szöveges mezők adattartalmát idézőjelek ("doublequote") közé kell tenni.

7. A karakterfelismerés (OCR) paramétereit:

7.1. Általános követelmények:

- A főszöveg karakterfelismerési pontossága minimum 95%, azaz 100 karakter maximum 5 hibát tartalmazhat.
- A digitalizált oldalak automatikus karakterfelismerést (OCR) követően kétrétegű PDF/A-1 (ISO 19005-1:2005) formátumban kerülnek elmentésre.
- A karakterfelismerés a mellékletek és a képek címére is terjedjen ki.
- Alkalmazott karakterkészlet: UNICODE karakterkészlet, UTF-8 kódolásban, normalizált formában, az ékezetes karakterek: „precombined” formában.
- A szkennelési hibák kézi javítása nem követelmény, az általános karakterfelismerés pontossági követelményén felül. Ezen kivétel azonban nem érintheti a speciális OCR követelmények betartását.

7.2. Speciális OCR követelmények:

Az Ajánlatkérő az általános OCR minőségi követelményekhez képest szigorúbb követelményrendszert állapít meg (azaz a karakterfelismerés 999 ezrelékes pontosságú kell legyen, vagyis 1000 karakterből 1 lehet hibás.), az alábbi dokumentum csoportok, illetve dokumentumrészletek esetében:

- Iromány: irományszám, iromány címe, benyújtás dátuma, benyújtó/k neve, digitalizált fájl neve.
- Bizottsági jegyzőkönyvek: bizottság neve, ülés dátuma, ülés száma, digitalizált fájl neve.

7.3. Az OCR pontosságának ellenőrzése:

Ajánlatkérő a meghatározott követelményeket véletlenszerű mintavétellel kívánja ellenőrizni.

8. Mentés, archiválás:

8.1. Fájlnev:

- Ajánlatkérő a képeket egyedi azonosító jelzettel kéri ellátni, ez a fájlnev. Az elnevezések kialakítása során Ajánlattevő egyeztetni köteles az Ajánlatkérővel. Az országgyűlési adatbázis számára elvárt filenév konvenciókat a 3. melléklet tartalmazza.
- A képfájlok elnevezése csak az angol ABC betűit, számokat, írásjelek közül pedig csak az alsó vonal jelet: _ (UTF-8 name: LOW LINE, code: U+005F), illetve a kötőjelet: - tartalmazhatja. Más karakter használata nem megengedett.

9. A digitalizált állományok átadása:

9.1. TIFF-állomány: Vállalkozónak a szkennelt állományokat veszteségmentesen tömörített TIFF formátumban, két példányban, 3,5”, minimum 1 TB SATA merevlemezen, USB 3.0 csatlakozásra képes, fémházas hordozható meghajtó egységen kell átadnia. A külső meghajtó egységen fel kell tüntetni a lemez tartalmát.

9.2. PDF-állomány: Vállalkozónak az optikai karakterfelismertetéssel (OCR) elkészített állományokat PDF/A-1 (ISO 19005-1:2005) formátumban, két példányban, 3,5”, minimum 1TB SATA merevlemezen, USB 3.0 csatlakozásra képes, fémházas hordozható meghajtó egységen kell átadnia. A külső meghajtó egységen fel kell tüntetni a lemez tartalmát.

9.3.A Vállalkozó által rendelkezésre bocsátott adathordozó a Megrendelő tulajdonává válik.

9.4.A külső meghajtó egység átadása során jegyzőkönyvet kell felvenni, amely tartalmazza:

- az átadás idejét,
- az átadás helyét,
- a résztvevők személyét, aláírását,
- az átadott külső meghajtó egység jellemzőit, mennyiségét, tartalmát,
- és a külső meghajtó egység működőképességét jellemző leírást.

9.5. Külső meghajtónként Ajánlatkérő kéri megjelölni és megadni:

- külső jelöléssel: meghajtó azonosítója/átadás dátuma. pl. DTT2-OGYK-01/20150321
- elektronikusan a gyökér könyvtárban: meghajtó azonosítója/átadás dátuma/dokumentum típusonként/dokumentumok darabszám
- CD-n és kinyomtatva is az átvételi listát, amely tartalmazza: meghajtó azonosítója/átadás dátuma/dokumentum típusonként/dokumentumok lemezre rögzített fájlok listája, amit átvételkor, de később is össze illetve be lehet azonosítani a lemez-tartalommal.

9.6. Az iratokat vonalkóddal ellátni nem lehet, csak a dobozokat.

9.7. Nem követelmény az elektronikus iratok archiválása.

- 9.8. A műszaki leírásban előírt hibahatár a dokumentumok összességére tekintettel értendő.
- 9.9. A dokumentumokat egy az egyben kell szkennelni és menteni. Ez vonatkozik-e a szövegtükörrre is.
- 9.10. A dokumentumok esetenként kézírást is tartalmaznak, amely az OCR karakterfelismerő program eredményességét is befolyásolja.
- 9.11. A Level A és a Level B formátumú pdf egyaránt elfogadható