

**PIM-DBK**

**dHUpla**

Born digital  
Definíciók, célok

[Jira DBT-9](#)

Kalcsó Gyula–Szűcs Kata

2021. január 18.

# Tartalom

## [Az adathordozók tipológiája](#)

### [Gyűjteményi adathordozók](#)

#### [Beépíthetők \(beépítve vagy külön\)](#)

#### [Külső eszközök](#)

### [Harmadlagos adatforrások](#)

## [A born digital objektumok tipológiája](#)

### [Fájltípusok](#)

### [Fájlok együttes kezelése](#)

## [A célok](#)

### [Digital Stewardship](#)

### [OAIS \(Open Archival Information System\) Reference Model \(1999\)](#)

### [Archiválás](#)

### [Kutathatóvá tétel](#)

#### [A szöveges tartalom és bizonyos metaadatok kinyerése \(Apache Tika\)](#)

### [Publikálás](#)

## [A dHUpla hatásköre](#)

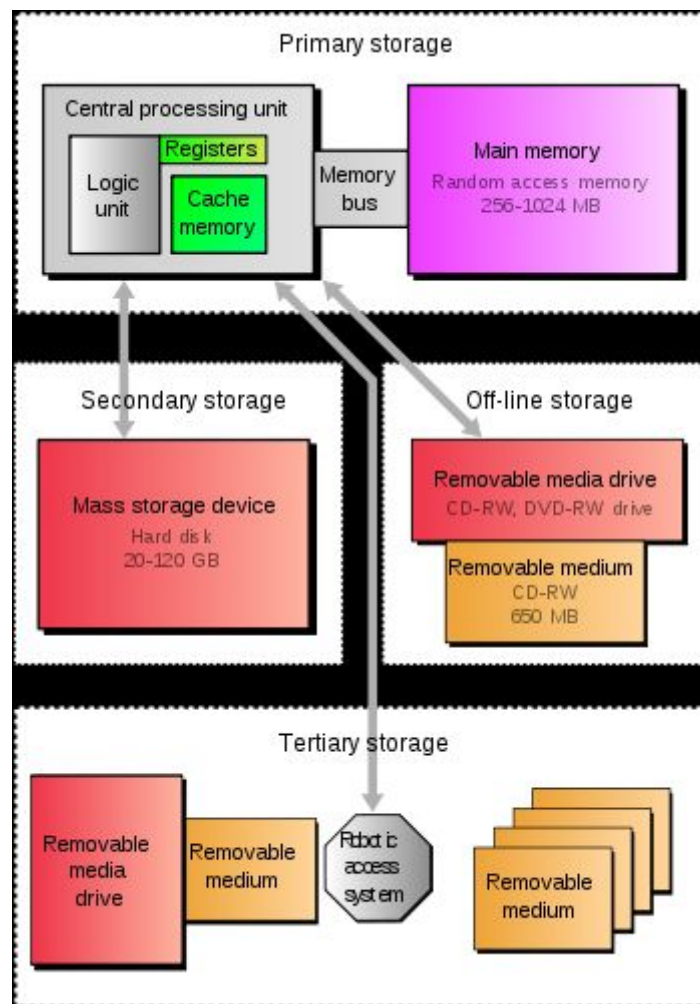
### [Az egyes born digital objektumtípusokra vonatkozó protokollok](#)

### [A DIP-csomag transzformációja a dHUpla rendszerében](#)

### [Az Apache Tika által támogatott formátumok](#)

# Az adathordozók tipológiája

Az adathordozók áttekintése azért szükséges, mert az egyes típusaik más-más kezelést, más workflow-t igényelnek. Megkülönböztethetők elsődleges, másodlagos és harmadlagos adathordozók.



Forrás: [https://en.wikipedia.org/wiki/Computer\\_data\\_storage](https://en.wikipedia.org/wiki/Computer_data_storage)

A dHUpa hatáskörébe a **on-line secondary** (ezeket fogom **beépíthető**nek nevezni) és az **off-line secondary** (ezeket fogom **külső v. leválasztható** eszközöknek nevezni) adathordozók, valamint a **tertiary** adathordozók közül a **felhőtárhelyek**, ill. egyéb, **távoli szerveren tárolt adatok** (pl. webes szolgáltatások, mint a közösségi médiaoldalak adatai) tartozhatnak.<sup>1</sup>

<sup>1</sup> Kérdés, hogy egy számítógép RAM-jából, ami elsődleges adathordozó (primary storage device), kiszedhető-e valami számunkra értékes információ, ill. hogy akarunk-e foglalkozni mobil eszközök memóriájában található adatokkal (az is elsődleges).

## Gyűjteményi adathordozók

Ezek olyan adathordozók, amelyek a PIM (vagy más intézmény) gyűjteményében fizikailag hozzáférhetőek (következésképpen csak elsődleges vagy másodlagos adathordozók lehetnek). Minden ide tartozó adathordozó esetében a FRED használatára lesz szükség.

### Beépíthetők (beépítve vagy külön)

HDD, SSD, mobil eszköz memóriája (?)

### Külső eszközök

Mágnesszalagok (tekercsen vagy kazettában – ezeken nem valószínű, hogy born digital tartalom van), hajlékonylemezek (8-, 5¼-, 3½-inch), memóriakártyák, pen drive-ok, optikai lemezek (CD, DVD, Blu-ray), esetleg más (?).



## Harmadlagos adatforrások

A felhőtárhelyről származó vagy valamilyen webes szolgáltató (pl. közösségimédia-oldal) által tárolt adatokat csak akkor tudjuk kezelni, ha azokról szabványos export (mentés) áll rendelkezésre. Meg kell határozni azokat a formátumokat, amelyeket be tudunk fogadni. A továbbiakban a kérdés a digitális objektumok tipológiájához, valamint a dHUpa hatásköréhez tartozik (l. következő rész).

# A born digital objektumok tipológiája

A born digital objektumok minden esetben fájlok. Egy objektum legalább egy, de sokszor több fájlból áll, amelyek lehetnek könyvtárakba szervezve.

## Fájltípusok

Az alábbi fő fájltypusok léteznek (vastaggal jelölve a digitális bölcsészeti szempontból releváns tartalmat hordozhatókat):

**Archívumok és tömörített fájlok** (ezen belül a lemezképek, de pl. ide tartoznak az archivált weboldalak becsomagolva is)

CAD-fájlok

**Adatbázisfájlok**

**Asztali kiadványszerkesztési formátumok**

**Dokumentumok**

Pénzügyi rekordfájlok

Fontfájlok

Geodatafájlok

**Grafikai formátumok (raszter- és vektorgrafika, 3D stb.)**

Linkfájlok

Matematikai fájlok

Futtathatók

**Nyomtatásikép-leírók (CSS, XSL is ide tartozik)**

**Személyes információkezelők (pl. PST)**

**Prezentációk**

Projektmenedzsment-kezelők

**Hivatkozáskezelők (pl. BibTeX)**

Tudományos adatfájlformátumok

Szkriptek

Biztonsági fájlok

Sugárzott adatformátumok

**Hangfájlok**

Lejátszási listák

Zeneszerkesztők fájljai

Televíziós felvételi formátumok

Számítógépes programok forráskódjai

**Táblázatok**

**Tabulált adatok (csv, tsv)**

**Videófájlok**

Videójáték-formátumok

Videójáték-ROM-formátumok

Virtuális gépek

**Webfájlok**

**Jelölőnyelvek fájljai (pl. RSS, EML, XML, JSON stb.)**

Egyebek

(Forrás: [https://en.wikipedia.org/wiki/List\\_of\\_file\\_formats](https://en.wikipedia.org/wiki/List_of_file_formats))

A nem szöveges tartalom esetében csak akkor lehet megfontolni a további feldolgozást, ha releváns, szöveggé alakítható tartalom található benne. (Pl. képfájlon OCR, hangzó szöveg átalakítása beszédfelismerővel stb.).

Bizonyos formátumok csak más fájltypussal együtt kezelendők (pl. egy XSL csak akkor érdekes, ha van hozzá XML, amit megjelenít).

## Fájlok együttes kezelése

Esetek, amikor pl. érdemes a fájlokat együtt tartani:

- Tanulmánygyűjtemény (minden tanulmány külön fájlban)
- Verseskötet (akár versenként, vagy ciklusonként külön fájl)
- Regény (fejezetenként külön fájlban)
- Levelezés
- Az adathordozón lévő tartalom részben vagy teljesen strukturált fájlhalmaz (pl. archivált weboldal)
- Stb.


A fájlok együttes kezelésére szolgáló eljárás lehet a Baglt.

### The Baglt file package format

John Kunze and Stephen Abrams, California Digital Library (CDL)


#### Replicas for a rainy day

- Need a data package format that can carry any data
- Suitable for disk-based and network-based transfer
- Main purpose is to move data from one digital library or archive to another for safe-keeping
- Not important whether the receiver provide access to or even understand meaning of the sender's content



**Fig. 1.** Sleeping better at night is more likely for the archivist who has replicas safely tucked away at other memory organizations.

1. start: put your files...
2. in a directory, data
3. create a tag
4. put tag next to data




**Bag&Tag**

#### Summary


Baglt is a file hierarchy, suitable for disk- or network-based transfer, and maybe bag return, with just enough structure to safely enclose a manifest, checksums, tag info, and an arbitrary payload.

#### Bags in action



**Library of Congress (LC) grantees send bags**

- For example, bagged web crawls funded by LC are sent by CDL to LC, with replica sent to San Diego



#### A Baglt bag directory listing

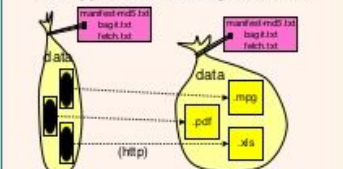
- A Baglt bag is a special directory (or Windows folder)
- The directory name is your choice, but inside the top level Baglt reserves names for required files:
  - data: a subdirectory where you can put anything
  - bagit.txt: a 2-line file declaring this is a bag
  - manifest-md5.txt: a list of files present

```
MyFirstBag/
├── manifest-md5.txt
├── bagit.txt
├── bag-info.txt
└── data/
    ├── my.pdf
    ├── my.xls
    └── my.mpg
```

**Fig. 2.** Payload directory, data, holds anything you want. "Tag" files (purple) describe the bag itself, including list of files and checksums in the manifest (algorithms other than md5 possible) and an optional "bag-info.txt" metadata file.

#### Holey bags! for efficiency

- A common optional tag file lists "holes" to be filled
- Payload is incomplete until these files are fetched
- File fetch.txt lists URLs for receiver to grab
- Benefits include (network transfers only):
  - No need to stage extra data copy at either end
  - Cheap parallelism from fetching URLs in batches



**Fig. 3.** On the left is a bag received with missing files, or "holes". The "fetch.txt" file lists URLs and corresponding payload filenames that the receiver must fetch before declaring the bag on the right complete.

#### Baglt credits and details

Authors from LC and CDL: Andy Boyko, John Kunze, Justin Littman, Liz Madden, Brian Vargas

Many thanks to: Stephen Abrams, Mike Ashenfelder, Scott Fisher, Erik Hetzner, Keith Johnson, David Loy, Tracy Seneca, Mark Phillips, Adam Turoff, Jim Turtle

**Baglt's specification:**

<http://www.cdlib.org/programs/cdl/infobag/bag-spec-0.3.txt>  
<http://www.cdlib.org/infobag/bag-spec.html>

Baglt was informed by:

- Enclose-and-Deposit method, Tabara & Sagimora,
- LC's eDeposit Pilot and NDIPP AHT
- ARC/WARC aggregate file format

#### For further information

Please contact [john.kunze@cdlib.org](mailto:john.kunze@cdlib.org) or [stephen.abrams@cdlib.org](mailto:stephen.abrams@cdlib.org)

For information on CDL's Preservation Program, see <http://www.cdlib.org/programs/cdl/preservation.html>

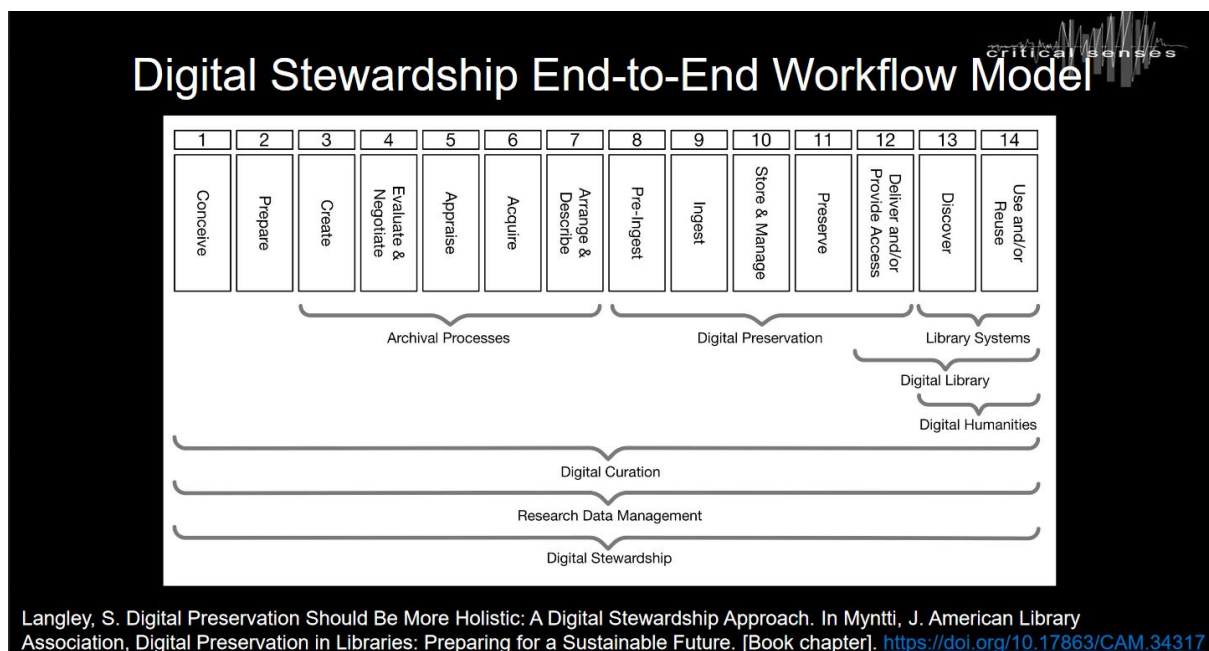
A Baglt formátumú csomagok szolgálhatnak az OAIS-modell szerinti SIP-, AIP- és DIP-csomagokként (l. lentebb).

# A célok

Érdeemes a PIM, a DBK és a dHUpa céljait különválasztani. A PIM szempontjából a muzeológiai feladatok megoldása a cél, azaz a már meglévő gyűjteményi born digital anyagot feltárni, gyűjteményezni, archiválni. A DBK feladata a PIM tevékenységének támogatásán túl ajánlások kidolgozása más közgyűjtemények részére, valamint a PIM-es és máshonnan érkező born digital anyagok dHUpába integrálásának a kidolgozása és elvégzése. A dHUpa a befogadott anyagokat szolgáltatja (kutathatóvá teszi és/vagy publikálja). (L. még [A dHUpa hatásköre](#) c. részt.)

Mivel a digitális anyagok kezelése (digital stewardship, digital curation, l. Langley 2019, 1. ábra) összetett folyamat, meg kell határozni, hogy a (1) PIM, a (2) DBK, valamint a (3) dHUpa mely fázisokkal foglalkozik.

## Digital Stewardship



A digitális anyagok kezelésének (digital stewardship, digital curation) teljes folyamata <https://www.repository.cam.ac.uk/bitstream/handle/1810/287006/EDITIONS%20ALCTS%20Digital%20%20CH%207.pdf?sequence=1&isAllowed=y>

Mivel a PIM-ben jelenleg is található born digital anyagok, ezért a múzeum feladata legalább a 8–12-ig terjed, ami a befogadás előkészítésétől (pre-ingest) a hozzáférés előkészítéséig (deliver and/or provide access) terjed. Esetleg lehet szó arról, hogy a PIM a legelső lépéstől foglalkozzon a folyamattal, hiszen tervezetten (?) hoz létre born digital tartalmat. Ez viszont nem (csak) a DBK hatásköre (?).

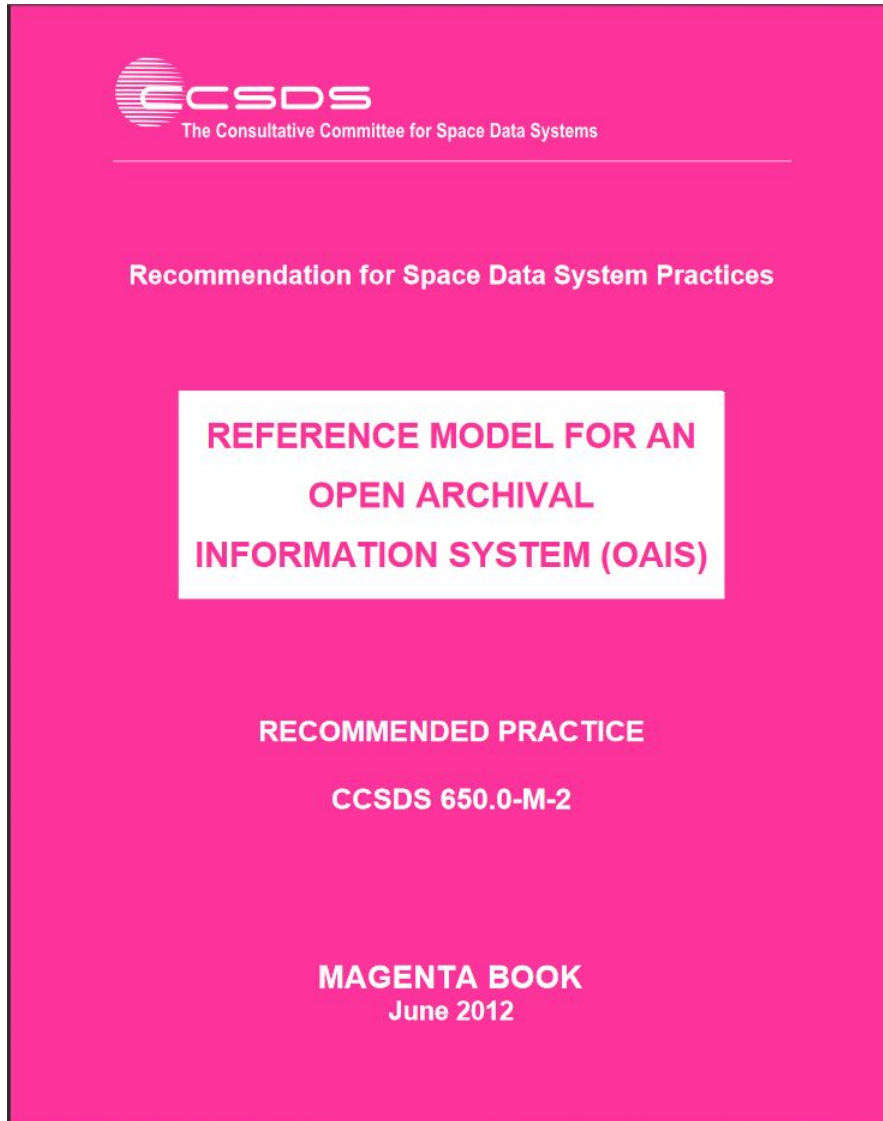
A DBK legalább a tárolás és kezelés (10. store & manage) fázisában be kell hogy kapcsolódjon, hiszen a dHUpa miatt a tárolt (archivált) objektumokkal is végez műveleteket, és mivel a szolgáltatás is a feladata, az egész folyamat a végéig a hatáskörébe tartozik.



Maga a dHUpla rendszere a bemenethez való hozzáférés biztosításától (12. deliver and/or access) a folyamat végéig tartó hatáskörrel bír.

Ugyanakkor a külső ajánlások miatt szükséges foglalkozni mindhárommal, tehát a 8–14-gyel (vagy ha a PIM az elejétől ellátja a feladatokat, akkor az egészszel).

## OAIS (Open Archival Information System) Reference Model (1999)

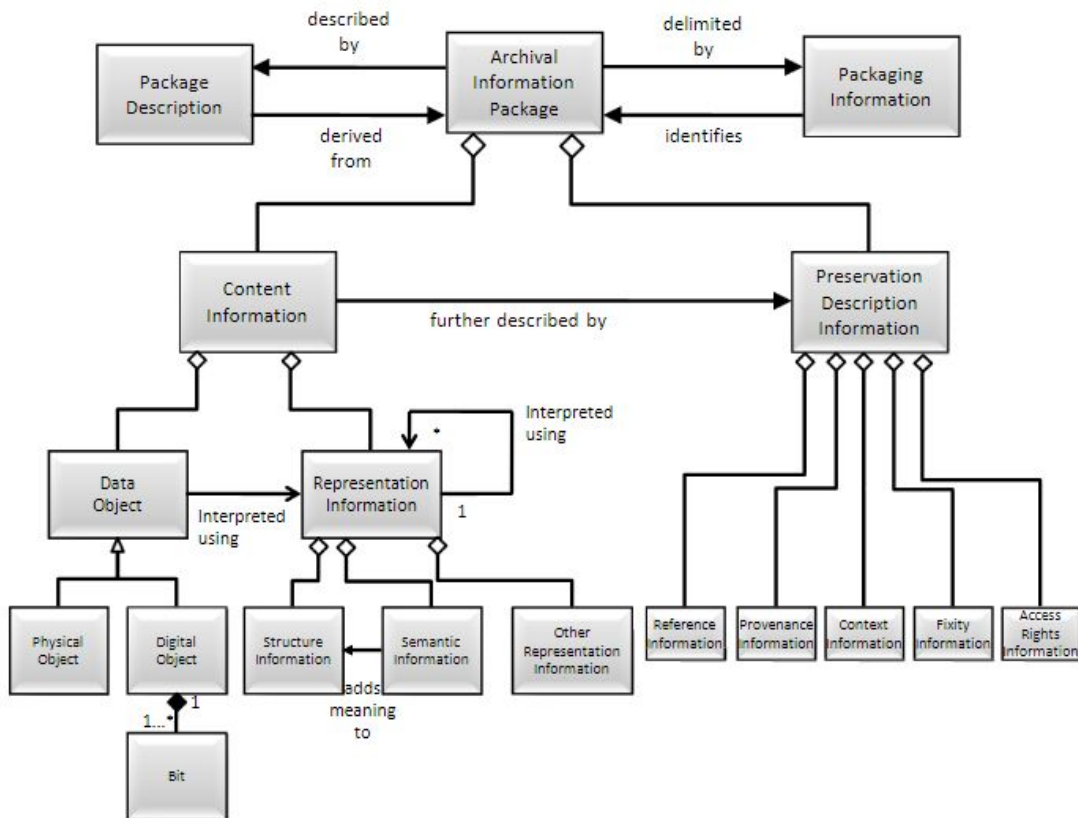
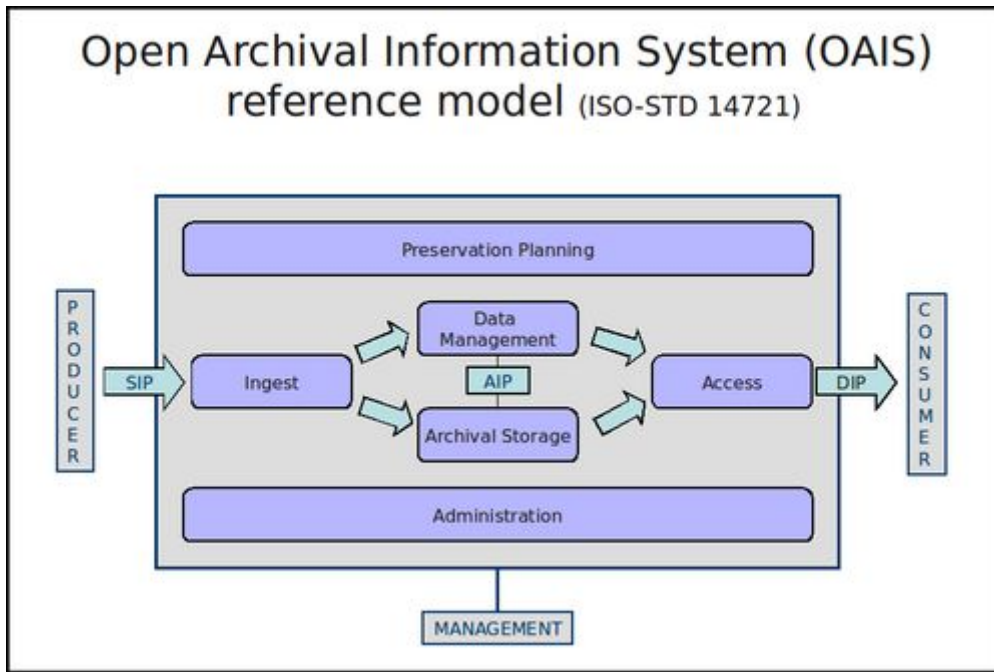


ISO standard (ISO 14721:2012). Ráadásul Magyarországon is van példa a használatára (MNL).

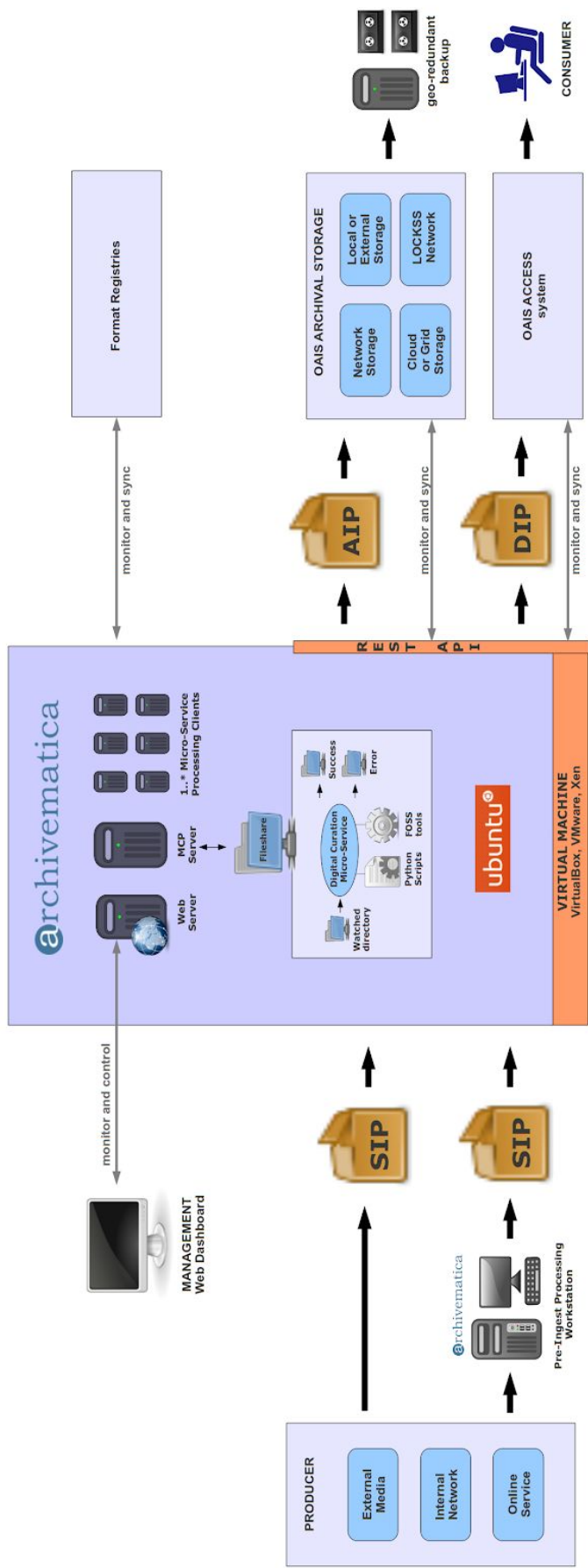
A modell digitális objektumokban gondolkodik, amelyek fájlokból és metaadatokból állnak. Több OS szoftver erre épül (Archivematica, Preservica, RODA, Rosetta)

Három csomagot (fájlok + metaadataik) ír elő:

- SIP (Submission Information Package),
- AIP (Archival Information Package),
- DIP (Dissemination Information Package).

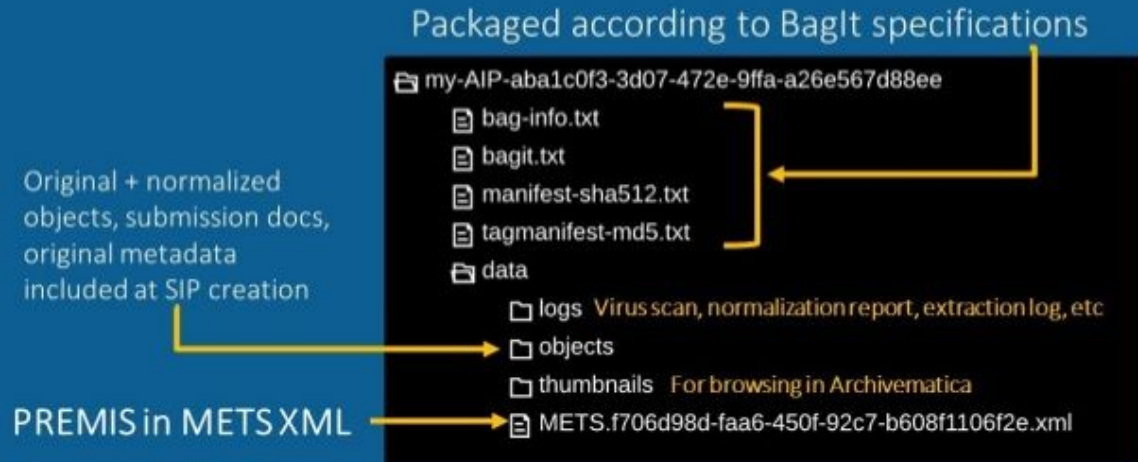


**Figure 4-18: Archival Information Package (Detailed View)**

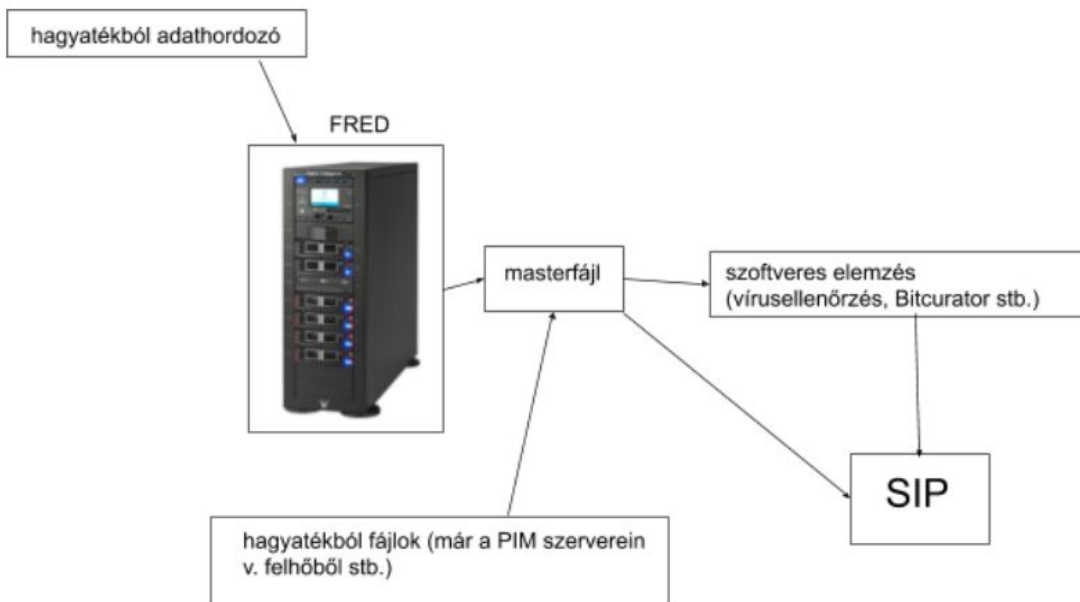


Az OAIS-modell csomagjai az Archivematicában  
 Forrás: [https://wiki.archivematica.org/Metadata\\_syncing](https://wiki.archivematica.org/Metadata_syncing)

# Archivematica AIP structure



Forrás: Slide Dan Gillean (Archivematica) prezentációjából, amelyet a Glenstone Digital Preservation Roundtable-ön tartott. Potomac, Maryland, 2016. november 4.



SIP-csomag előállítására gyűjteményi adathordozón található anyagból

A csomagok tartalmára vonatkozóan léteznek részletes ajánlások: <https://dilcis.eu/>.

## Archiválás

Az archiválás három útja: (1) PIM saját infrastruktúrája, (2) OS szoftverek, (3) felhőszolgáltatás.

Az (1) adott, a fejlesztésén érdemes gondolkodni.

Az OS-szoftverek használatának korlátja lehet, hogy többnyire linuxos platformfüggésük van. Pedig nagyon jó, széles körben elterjedt, az OAIS-modellnek is megfelelő szoftverek vannak: Archivematica, ArchiveSpace (felhőt is tud), RODA (ez utóbbi tűnik talán a legjobbnak, EU-s támogatást is élvez).

A felhőszolgáltatást is nyújtó archívumok esetében a platformfüggetlenség nem gond, viszont a tárhely miatt nem ingyenesek (vagy nagyon korlátozottak): ExLibris Rosetta, Preservica.

Bármelyiket is választjuk, a born digital objektumok esetében AIP-csomagban kell archiválni.

## Kutathatóvá tétel

Adathordozók és lemezképek esetén maga a struktúra is érdekes lehet, azt is érdemes kutathatóvá tenni. Ezek esetében szükséges egy olyan DIP-csomag előállítás, amelyben a rejtett fájlok is benne vannak.

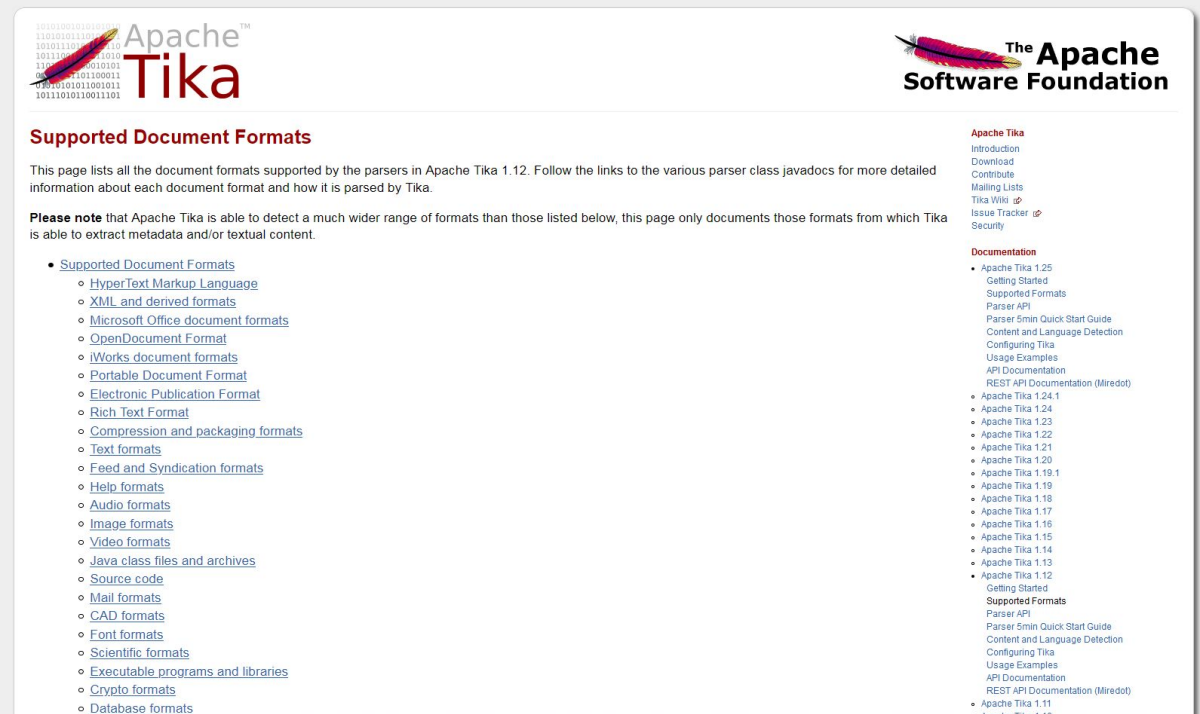
Az egyes fájlok (szöveges) tartalmának a licensze kérdés (csak megtekinthető v. valahogyan exportálható, letölthető?).

Az egész folyamat (csakúgy, mint a dHUpába való befogadás, l. alább) függ a típustól, azaz külön protokollok szükségesek.

Kérdés, hogy milyen formátum(ok) szolgál(nak) a kutathatóvá tételre.

## A szöveges tartalom és bizonyos metaadatok kinyerése (Apache Tika)

Erre van eszköz: [Apache Tika](#). Google [Tesseract](#)tal képes OCR-re is, ill. hangfelismerést is tud (tesztelni kell magyarra). Ha nem is ezt használjuk, de ilyenre szükség van.



**Supported Document Formats**

This page lists all the document formats supported by the parsers in Apache Tika 1.12. Follow the links to the various parser class javadocs for more detailed information about each document format and how it is parsed by Tika.

**Please note** that Apache Tika is able to detect a much wider range of formats than those listed below, this page only documents those formats from which Tika is able to extract metadata and/or textual content.

- [Supported Document Formats](#)
  - [HyperText Markup Language](#)
  - [XML and derived formats](#)
  - [Microsoft Office document formats](#)
  - [OpenDocument Format](#)
  - [iWorks document formats](#)
  - [Portable Document Format](#)
  - [Electronic Publication Format](#)
  - [Rich Text Format](#)
  - [Compression and packaging formats](#)
  - [Text formats](#)
  - [Feed and Syndication formats](#)
  - [Help formats](#)
  - [Audio formats](#)
  - [Image formats](#)
  - [Video formats](#)
  - [Java class files and archives](#)
  - [Source code](#)
  - [Mail formats](#)
  - [CAD formats](#)
  - [Font formats](#)
  - [Scientific formats](#)
  - [Executable programs and libraries](#)
  - [Crypto formats](#)
  - [Database formats](#)

**Apache Tika**  
Introduction  
Download  
Contribute  
Mailing Lists  
Tika Wiki  
Issue Tracker  
Security

**Documentation**

- [Apache Tika 1.25](#)
  - [Getting Started](#)
  - [Supported Formats](#)
  - [Parser API](#)
  - [Parser 5min Quick Start Guide](#)
  - [Content and Language Detection](#)
  - [Configuring Tika](#)
  - [Usage Examples](#)
  - [API Documentation](#)
  - [REST API Documentation \(Miredot\)](#)
- [Apache Tika 1.24.1](#)
- [Apache Tika 1.24](#)
- [Apache Tika 1.23](#)
- [Apache Tika 1.22](#)
- [Apache Tika 1.21](#)
- [Apache Tika 1.20](#)
- [Apache Tika 1.19.1](#)
- [Apache Tika 1.19](#)
- [Apache Tika 1.18](#)
- [Apache Tika 1.17](#)
- [Apache Tika 1.16](#)
- [Apache Tika 1.15](#)
- [Apache Tika 1.14](#)
- [Apache Tika 1.13](#)
- [Apache Tika 1.12](#)
  - [Getting Started](#)
  - [Supported Formats](#)
  - [Parser API](#)
  - [Parser 5min Quick Start Guide](#)
  - [Content and Language Detection](#)
  - [Configuring Tika](#)
  - [Usage Examples](#)
  - [API Documentation](#)
  - [REST API Documentation \(Miredot\)](#)
- [Apache Tika 1.11](#)
- [Apache Tika 1.10](#)

A Tika parser outputja plain text, html vagy xhtml. A metaadatok formátuma XMP (RDF-XML).

## Publikálás

Csakúgy, mint a dHUplába való befogadás (l. alább), függ a típustól, azaz külön protokollok szükségesek.

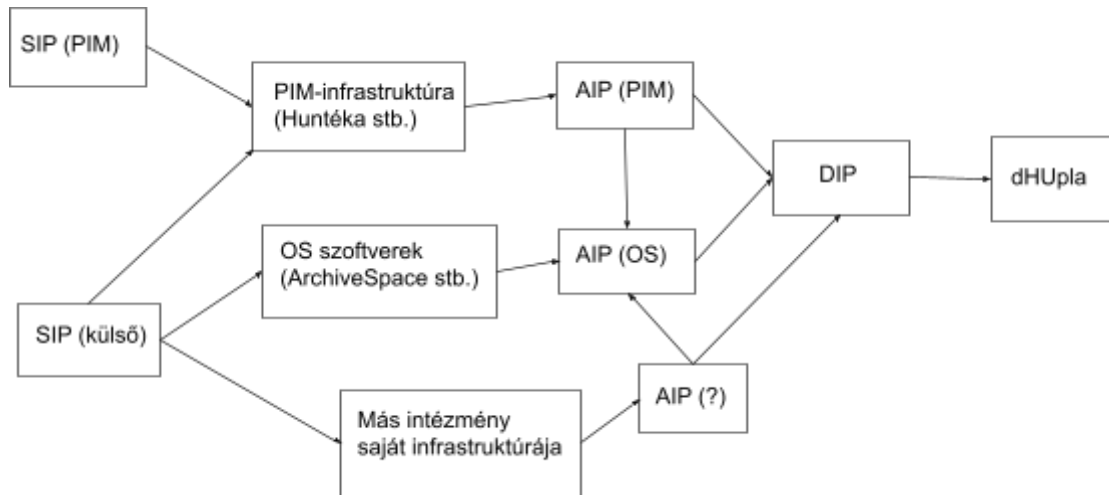
Kérdés, hogy milyen formátum(ok) szolgál(nak) a publikálásra.

Végső soron megjeleníteni fájlokat lehet. Van-e szükség másra, mint a repozitóriumban való megjelenítés? A kérdés az, hogy az összefüggő struktúrákat (pl. hypertext) hogyan kezeljük.

A born digital szöveges tartalmakból is célszerű TEI-XML-t csinálni, innentől pedig a megjelenítés problémái ugyanazok, mint a forráskiadás vagy a kritikai kiadás esetében.

# A dHUpla hatásköre

A készülő platform célja kettős: egyfelől szükséges a PIM meglévő és jövőbeli born digital anyagainak a kezelése, másfelől ajánlást dolgoz ki a magyar GLAM-szektor számára bizonyos típusú born digital objektumok kezelésére. A fejezet célja, hogy meghatározza a feldolgozható digitális objektumok körét.



A dHUplába olyan born digital objektumokból készült csomagok kerülhetnek be, amelyek megfelelnek az előre meghatározott, DIP-csomaggal szemben támasztott elvárásoknak.

## Az egyes born digital objektumtípusokra vonatkozó protokollok

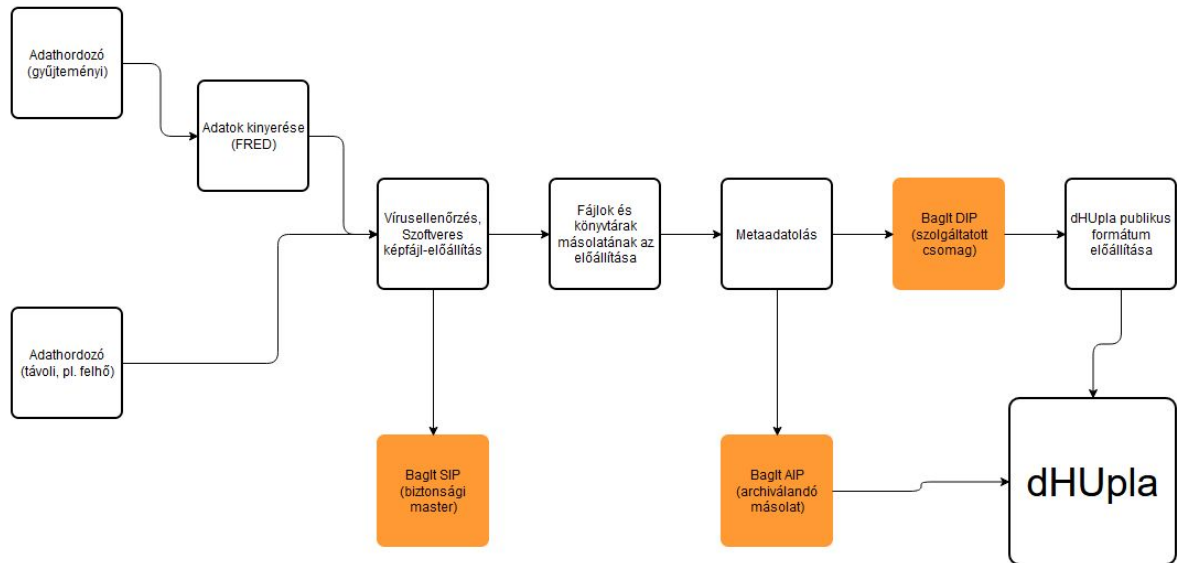
külső adathordozó (részben) összefüggő tartalommal → strukturált fájlhalmaz csomagban (pl. e-mail levelezés)



külső adathordozó össze nem függő fájlokkal → fájl (pl. egy különálló PDF, egy prezentáció stb.)

Ugyanakkor érdemes arra is gondolni, hogy nem lesz minden automatizálható, és az egyedi esetekre is ki kell dolgozni valamilyen protokollt.





## A DIP-csomag transzformációja a dHUpla rendszerében

A következő gyakorlati kérdés az, hogy milyen fájlokból lehet szolgáltatott formátumot, pl. TEI-XML-t csinálni (lehetőleg automatikusan). Az Apache Tika által előállított .txt vagy .xhtml pl. lehet egy opció. A nagy kérdés az, hogy ki és hogyan fogja elvégezni a szövegek tisztítását.

Mivel az Apache Tika bemenete lehet (akár más formátumból, pl. JSON-ból előállított) XML, ezért akár a közösségimédia-exportokat is tudja kezelni.

# Az Apache Tika által támogatott formátumok

## File Formats Supported by Tika

The following table shows the file formats Tika supports.

File format	Package Library	Class in Tika
XML	org.apache.tika.parser.xml	XMLParser
HTML	org.apache.tika.parser.html and it uses Tagsoup Library	HtmlParser
MS-Office compound document Ole2 till 2007 ooxml 2007 onwards	org.apache.tika.parser.microsoft org.apache.tika.parser.microsoft.ooxml and it uses Apache Poi library	OfficeParser(ole2) OOXMLParser (ooxml)
OpenDocument Format openoffice	org.apache.tika.parser.odf	OpenOfficeParser
portable Document Format(PDF)	org.apache.tika.parser.pdf and this package uses Apache PdfBox library	PDFParser
Electronic Publication Format (digital books)	org.apache.tika.parser.epub	EpubParser
Rich Text format	org.apache.tika.parser.rtf	RTFParser
Compression and packaging formats	org.apache.tika.parser.pkg and this package uses Common compress library	PackageParser and CompressorParser and its sub-classes
Text format	org.apache.tika.parser.txt	TXTParser
Feed and syndication formats	org.apache.tika.parser.feed	FeedParser
Audio formats	org.apache.tika.parser.audio and org.apache.tika.parser.mp3	AudioParser MidiParser Mp3- for mp3parser
Imageparsers	org.apache.tika.parser.jpeg	JpegParser-for jpeg images
Videoformats	org.apache.tika.parser.mp4 and org.apache.tika.parser.video this parser internally uses Simple Algorithm to parse flash video formats	Mp4parser FlvParser
java class files and jar files	org.apache.tika.parser.asm	ClassParser CompressorParser
Mobxformat (email messages)	org.apache.tika.parser.mbox	MobXParser
Cad formats	org.apache.tika.parser.dwg	DWGPParser
FontFormats	org.apache.tika.parser.font	TrueTypeParser
executable programs and libraries	org.apache.tika.parser.executable	ExecutableParser