# DataModelUseCases

## Data Model Use Cases

## PDF document

Still probably the single most widely used use case. An end user uploads a PDF and appropriate metadata. The system extracts the full-text automatically.

Later, an updated PDF is uploaded, because the original PDF contained some spelling mistakes.
Published article

The preprint of a journal article has been deposited in an archive for Open Access. Later the published article is uploaded, with an embargo period of 6 months. The published article is created as a new Item, with its relationship to the preprint indicated in the metadata.

## Photo

A photo is uploaded as a high-resolution TIFF file. Two lower-resolution JPEG files are created - one suitable for a 'preview' display in a typical Web browser, and the other a small thumbnail for use in lists of search results and the like.

## Archived Web site

The snapshot of a Web site at a particular point in time is uploaded. It contains tens of files and several different file formats. It may contain links to external Web sites, i.e. pages that are not inside the archived item. The HTML may function correctly only in a particular browser (e.g. IE).

## Word document with PDF and XML/HTML versions

A source Word document (potentially with embedded images, and perhaps the original "source" files for those images (e.g. Visio files) are uploaded to a DSpace instance. From this document, a PDF/A version is produced via some conversion tool, as is an HTML version, consisting of HTML, GIF and CSS format files. The full text of the Word document is also extracted for full-text indexing.

The source Word document is later modified to fix some typos. This means that the PDF/A version, HTML version, and extracted full text must be re-converted.

## Book

A set of scanned pages, at high resolution and lower resolution. OCR'd text. PDF which has the scanned images with OCR'd text "embedded".

## Resume/CV

Word doc with PDF version?

## Thesis

## Video

A training video in the AVI format is uploaded. The system needs to identify and store which encodings are used for both the audio and video streams. A lower-resolution copy for streaming may be required. Close caption text may be extracted and stored too.

## Dataset

Big XML file or spreadsheet, codebook, potentially an XML Schema

## Archive (e.g. .zip)

How to store the formats of the contained files

## Extracting metadata

PDF, JPEG, other formats have embedded metadata

## Anthropological study

Notes, audio interview, couple of videos – issue is it necessary to accurately work out what are the atomic 'representations'?

## Software

(Stretch)