PDF - Text extraction Specs

Example

Example

```
<?xml version="1.0" encoding="UTF-8"?>
<nitf xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://iptc.org/std/NITF/2006-10-18/ nitf-3-6.xsd"
xmlns="http://iptc.org/std/NITF/2006-10-18/" version="-//IPTC//DTD NITF
3.6//EN">
    <head>
        <title>Title</title>
    </head>
    <body>
        <body.head>
            <hedline>
                <hl2>Overtitle</hl2>
                <hl1>Title</hl1>
                <hl2>Subtitle</hl2>
            </hedline>
            <byline>Author</byline>
        </body.head>
        <body.content>
            Body
        </body.content>
    </body>
</nitf>
```

Article structure

- **Title**: Text located in the `<title>` tag in the `<head>` section and in the `<hl1>` tag in the `<hedline>` section
- **Overtitle**: Text located in the `<hl2>` tag before the title in the `<hedline>` section (optional)
- **Subtitle**: Text located in the `<hl2>` tag after the title in the `<hedline>` section (optional)
- **Author**: Text located in the `<byline>` tag in the `<body.head>` section (optional)
- **Body**: Text or HTML located in the `<body.content>` tag

Articles organization

Each article should have its own xml file, called art_XX.xml, were XX is the progressive article index. E.g. art_01.xml, art_02.xml ecc.

All the images used by the html img tags should published on a public URL.

The issue should be packed as a zip file with the following folder structure:

- <root>
    - issue.pdf
    - articles (folder)
        - article_01.xml
        - article_02.xml
        - article_03.xml
        - ....