

# eClips Full-Page Service

---

## Technical Documentation

*Version 2.1*

### Contents

Overview .....	2
Accessing the service .....	2
About the Content .....	2
Editions – Null and Regional .....	3
File names .....	3
Availability schedule .....	4
PDF files .....	4
XML files .....	5
Support and troubleshooting .....	6
Appendix 1: Sample data .....	7
Appendix 2: Titles covered .....	11

---

## Overview

The NLA Full Page Service provides a PDF for each page of a newspaper, and XML containing all text automatically extracted from that page.

Typically, the recipient will search the XML files to identify which pages are of interest, and then produce their own digital cuttings of relevant articles from the pages for their clients.

This document is designed for technical users to assist in planning and developing services based around the Full Page service. For commercial issues regarding usage, please refer to the licence agreement.

## Accessing the service

When customers subscribe to the eClips Full Page service they will be given login details by the NLA Client Services team. If there is any issues or changes required please contact NLA Client Services at [clientservices@nla.co.uk](mailto:clientservices@nla.co.uk)

The service provides PDF and XML files that are distributed using FTP. The FTP server can be accessed at <ftp://pagefeeds.nla-eclips.com> Access requires the use of the username and password provided by Client Services.

Please note that Active mode FTP is not currently supported and all connections should be in **Passive** mode.

## About the Content

The Full Page service delivers PDF and XML files for each page published in all titles covered (Appendix 2). As well as the main book of the newspaper ('null'); the service also delivers changed pages for regional variations, supplements, classified ad pages and can include cover-wraps.

Where further editions are produced to the 'null' edition (either regional variation, or time-based editions), all pages produced by the publisher for the new edition are included. If pages are updated before printing (e.g. as a result of a small amendment prior to print) then the original file will be overwritten, so only the pages used in the print run of each edition is included.

The text of each page is extracted automatically, and divided into paragraphs where possible. This automation of text extraction means that:

- The XML file will contain editorial and non-editorial text
- Text from Ads should be expected.
- There should be no expectation that the text will be ordered into articles
- Continuations (stories carried across multiple pages) are not handled
  - o where this occurs the customer will need to identify the subsequent pages.

---

## Editions – Null and Regional

The Full Page Service contains all pages provided to us by the publisher. In many cases they produce their titles based on 'changed pages'. This means that where there are multiple editions for a single title, or multiple regional variations of a single title, there will be:

A core edition, containing all pages (we refer to this as the 'null' edition)

For all other editions, only pages which are different to those in the core edition

To recreate the entirety of all editions the logic is:

Identify how many editions there are for a given publication – either use a pre-made lookup table, or programmatically assess the feed each day to identify the number of regional variations

Create one copy of the entire null edition for each and every additional region, amending the region identifier in the filenames as appropriate

Taking one entire copy at a time, overwrite any relevant pages with the "changed pages" supplied by the publisher

Example

Burnley Express comprises a null edition and a Padiham edition (changed pages edition).

Make 2 copies of the null edition

The first copy needs no amendment and is thus the entirety of the Burnley Express 'home' edition

Overwrite any changed pages in the 2<sup>nd</sup> copy, thus producing an exact replication of the Burnley Express Padiham edition

## File names

All files are saved into a single root folder, where they are uniquely and consistently named.

The naming format is Date\_Title\_Region\_Supplement\_Edition\_PageNumber, followed by the PDF or XML file extension. For example:

**20110408\_Nelson Leader\_Reg-null\_Sup-null\_Ed-01\_039.pdf**

**20110408\_Nelson Leader\_Reg-null\_Sup-null\_Ed-01\_039.xml**

Date	Year, Month, Day - yyyymmdd
Title	The title of the newspaper
Region	If there is no regional edition then Reg-null
Supplement	The name of the supplement. If it is the mainbook then Supp-null
Edition	1 <sup>st</sup> edition = Ed-01, 2 <sup>nd</sup> edition = Ed-02, etc.
PageNumber	nnn

---

## Availability schedule

New content will first start to become available in the FTP area at 10pm the day prior to publication, e.g. content for the papers on Tuesday will first appear in the FTP from 10pm on Monday.

Your delivery of Links to Clients and access to those links by Clients is **embargoed** till 4am.

PDF page files are posted first, with the XML for each page following once it has been processed, typically 1-2 minutes later.

Pages are delivered in a continual near-to real-time feed therefore a Schedule will be provided to give an indication of expected completion times for each edition.

As a guide, morning titles are usually complete and available the night before publication or early in the morning on the day of publication, and evening publications may be partially available by the morning of publication, but often complete later during the day. Therefore, it is recommended to poll the FTP site more than once during each 24-hour period, for example at 3am, 10am, and then at 1pm midday.

Pages are not delivered in consecutive pagination order for each edition, but are delivered in the order that they are received from the publisher and processed by the NLA.

Content will remain available on the FTP server for **2 days** following the date of publication (or the date the file was first made available via the FTP area), at which point it will be permanently purged.

## PDF files

The PDF files are as-sent to the publisher's print sites, with the exception that all pages pass through a process at the NLA to reduce the resolution of the image layer to 72dpi resolution. Text layer remains at the native resolution, typically 150 dpi, so any Optical Character Recognition (OCR) processes that a user wishes to perform on the PDF are unaffected.

The average PDF page size is 500 KB, but varies dependent on the proportion of text to images on the page. A full-page and complex image can be many MB whilst a text heavy page 200-300 KB.

Pages are either single, or double-page spreads. Note this can include 'printer's pairs', i.e. double-pages which are non-consecutive.

## XML files

Each XML file contains data extracted automatically from the PDF page(s). The structure of each file is consistent.

Element	Field tag	Notes
Edition	<b>Property FormalName="Edition"</b>	Some newspapers release multiple editions during the day. Numerical value.
Region	<b>Property FormalName="Region"</b>	Some newspapers are released under the same title in multiple regions. Free text value. If there is no regional variation, the value will be "null".
Supplement	<b>Property FormalName="Supplement"</b>	Some newspapers include additional supplements to the main paper (mainbook). Mainbook has value of "null". Other supplements are identified by a free text field according to their name.
Publication Date	<b>DateLineDate</b>	yyyymmdd, e.g. 17 <sup>th</sup> March 2011 is represented as 20110317.
Publication Name	<b>Property FormalName="Publication_Name"</b>	The title of the newspaper. Free text field.
Publication Acronym	<b>Property FormalName="Publication_Acronym"</b>	Each title is identified by the NLA with a unique acronym, e.g. JPHM.
Page Numbers	<b>Property FormalName="Page_Numbers"</b>	Numerical value. Multiple pages are delimited by commas. E.g. "001,002"
Page Type	<b>Property FormalName="Page_Type"</b>	Pages can either be "Single" or "Spread"
Issue	<b>Property FormalName="Issue"</b>	Same as Edition. This is retained for future use
Day	<b>Property FormalName="Day"</b>	Freetext for the publication day, e.g. "Thursday"
PDF reference	<b>PagePDF Href</b>	NLA use only. Filename for the publisher source PDF associated with this page.
Creation Time	<b>CreationTime</b>	Time at which this XML was generated.
File Size	<b>SizeInBytes</b>	Ignore. This will not be populated for Full Page delivery.
Height	<b>Property FormalName="Height"</b>	Height of the page. Supplied as a numerical value, along with an additional identifier for the units used, e.g. <b>Property FormalName="Height" Value="344" ValueRef="mm" /&gt;</b>
Width	<b>Property FormalName="Width"</b>	Width of the page. Supplied as a numerical value, along with an additional identifier for the units used, e.g. <b>&lt;Property FormalName="Width" Value="285" ValueRef="mm" /&gt;</b>
Structure	<b>Property FormalName="Structure"</b>	Ignore. This is retained for future use.
Text	<b>Text</b>	Paragraphs, where identified, are divided with <P>...</P>.

---

## Support and troubleshooting

In the event of any problems, please contact NLA Client Services during business hours at [clientservices@nla.co.uk](mailto:clientservices@nla.co.uk).



---

## XML

```
<?xml version="1.0" encoding="UTF-8" ?>
- <Data>
  <DateAndTime>20110324T172953+0000</DateAndTime>
  <DateId>20110317</DateId>
  <NewsItemId>thestar_20110317_002_city_01_null.pdf</NewsItemId>
- <DescriptiveMetadata>
  <Property FormalName="Edition" Value="01" />
  <Property FormalName="Region" Value="city" />
  <Property FormalName="Supplement" Value="null" />
  <DateLineDate>20110317</DateLineDate>
  <Property FormalName="Publication_Name" Value="The Star" />
  <Property FormalName="Publication_Acronym" Value="TSR" />
  <Property FormalName="Page_Numbers" Value="002" />
  <Property FormalName="Page_Type" Value="Single" />
  <Property FormalName="Issue" Value="01" />
  <Property FormalName="Day" Value="Thursday" />
  </DescriptiveMetadata>
- <PagePDF Href="thestar_20110317_002_city_01_null.pdf">
  <SizeInBytes />
  <CreationTime>20110324T172953+0000</CreationTime>
  <Property FormalName="Height" Value="344" ValueRef="mm" />
  <Property FormalName="Width" Value="285" ValueRef="mm" />
  <Property FormalName="Structure" Value="no" />
  </PagePDF>
  <PageThumbnail Href="thestar_20110317_002_city_01_null.jpg" />
- <PageContents>
- <Text>
  <P>2</P>
  <P>The Star, Thursday, March 17, 2011</P>
  <P>www.thestar.co.uk</P>
  <P>The Star, York Street, Sheffield, S1 1PU www.thestar.co.uk</P>
  <P>To enquire about permission to copy cuttings for internal management and information purposes please contact the NLA, Wellington Gate, 7 & 9 Church Road, Tunbridge Wells TN1 1NL. Telephone 01892 525273. Email copy@nla.co.uk</P>
  <P>The Star is printed on paper from recycled and sustainable sources. Please help the environment by recycling this issue. Recycled paper made up 87.2% of the raw material for UK newspapers in 2008.</P>
  <P>Weather</P>
  <P>TONIGHT</P>
  <P>CONTACT US</P>
  <P>TOMORROW</P>
  <P>THE WORLD</P>
  <P>DATA PROTECTION</P>
  <P>THE WEEK AHEAD</P>
  <P>YORKSHIRE: After a cold and frosty start there will be sunny spells, but also some wintry showers. Gentle winds. Max temp 6-9C (43-48F).</P>
  <P>By supplying your contact details, including email address and mobile number, you agree that Johnston Press plc, publishers of The Star and its business partners may contact you about new promotions, products and service by mail, email, phone, fax, SMS/MMS. Add the word STOP at the end of your communication if you do not wish to receive these. For quality and training purposes we may monitor communications. Service provided by JMedia UK Ltd, 0844 8001188. By submitting any contribution, you expressly grant Johnston Press Group plc a royalty-free licence to use such content in accordance with our terms and conditions at http://ww1.investorrelations.co.uk/jpplc/termsofaccess/ If you do not consent to this, you should not submit your contribution</P>
  <P>SATURDAY</P>
  <P>Dry with sunshine but also some cloud.</P>
  <P>SUNDAY</P>
  <P>Dry and fine with plenty of sunshine.</P>
  <P>MONDAY</P>
  <P>A fine and dry day with lots of sunshine.</P>
  <P>Country City.....Weather F C £ abroad Britain Sheffield..... fair 46 8 – Australia Sydney..... fair 75 24 1.53 Dlr Canada Toronto..... sunny 41 5 1.50 Dlr Canary Islands ..... cloudy 64 18 1.10 Euro Cyprus Larnaca..... sunny 68 20 1.10 Euro Denmark Cop'hagen
```



---

..... sunny 37 3 8.17 Kr France Paris ..... sunny 59 15 1.10 Euro Germany Berlin .....  
rain 37 3 1.10 Euro Greece Athens..... sunny 63 17 1.10 Euro Holland Amsterdam ..... fair 48  
9 1.10 Euro Ibiza San Antonio ..... cloudy 57 14 1.10 Euro Italy Rome ..... rain 57 14  
1.10 Euro Majorca Palma ..... cloudy 59 15 1.10 Euro Malta Valletta ..... sunny 66 19  
1.10 Euro Portugal Faro ..... fair 63 17 1.10 Euro Spain Barcelona ..... cloudy 57 14 1.10  
Euro Sth Africa Cape Town .. sunny 79 26 10.47 Rnd Switzerland Geneva .... fair 59 15 1.41 Fr  
Turkey Istanbul..... sunny 57 14 2.39 Turkey USA Miami ..... cloudy 79 26 1.54 Dlr  
USA New York..... cloudy 43 6 1.54 Dlr</P>

<P>SUN & MOON:</P>

<P>Moon: Rises 15.31 Sets 04.52</P>

<P>Sun: Rises 06.16 Sets 18.12</P>

<P>YORKSHIRE: Evening cloud and any rain will clear to leave a dry night with long clear spells, but also some cloud towards dawn. Cold and frosty. Gentle northwesterly winds. Min temp -2 to 1C (28-34F).</P>

<P>BRITAIN: A few wintry showers for north-west Scotland. The rest of Scotland along with northern England, the north Midlands and north Wales will become dry with long clear spells. It could become cloudier across north Wales, the north Midlands and Lancashire towards dawn. Cold and frosty across Scotland, northern England, north Wales and the north Midlands. Rather cloudy elsewhere with drizzle.</P>

<P>5</P>

<P>5</P>

<P>Call 0114 276 7676 News: extension 3510 Sport: ext 3344 Features: ext 4558 Photographic: ext 3342 Advertising: 0114 276 6666 Display: 0114 289 4020 Family notices: 0114 252 1207</P>

<P>LIGHTING UP:</P>

<P>London 18.08 - 06.09 Glasgow 18.23 - 06.25</P>

<P>Sheffield 18.12 - 06.14</P>

<P>£10k reward for pair who tackled gunman</P>

<P>Continued from page one</P>

<P>calm and wanted to keep everyone safe. "It was only afterwards when I thought about it that I realised I could have been shot - I was just thinking about everyone else." He admitted: "I was scared when the gunman started firing. A lot of people froze and others were running around and jumping, but Shanwaz got hold of the gun and I got the guy onto the floor and jumped on top of him. "He managed to shoot a few more times until we got the gun out of his hands, but we were pushing it towards the floor to stop him hitting anyone."</P>

<P>Saif added: "I am surprised the reward has been offered - it is really generous." Shanwaz, a taxi driver from Rotherham, said: "My sixth sense just kicked in when he pulled out the gun. I was close by so I just grabbed it without thinking.</P>

<P>"We might have been the first two there to disarm him, but other friends then got involved too and played their part. "A family occasion, a wedding, was ruined by this, and I am just sorry for the bride and groom and their parents that this happened. "My only thought at the time was to get the gun off the guy." The businessman said he was moved to offer the reward as he owes Saif and Shanwaz "everything". "I am a businessman in a fortunate enough position to be able to offer this reward," he said. "They put their lives on</P>

<P>the line for everyone at the wedding and this is my way of saying thank you to them. "All I know is a man walked in, pulled a gun out and started shooting. "I just froze with fear - I didn't know what was going on or what to do. "The next thing I saw was the two guys going over to him and one of them grabbing the gun from his hand and the other one hitting him and getting him to the floor, then wrestling with him until they got the gun off him." He said: "One of them put his hand over the barrel of the gun - if it had been fired again his hand would have</P>

<P>been blown off. It was terrifying. "The reason I want to give them the money is those boys are heroes in my eyes - I was standing close by and could easily have been shot, but I could not move because I was shaking. "My wife and kids were at the wedding, and they, too, could easily have been killed. "Everyone at that wedding must feel they owe their lives to Saif and Shanwaz." ■ A man has been charged with possession of a gun with intent to cause fear of violence, and two other men have been bailed pending further inquiries.</P>

<P>A HEAVY police presence has been placed around a Sheffield secondary school following a series of violent incidents between pupils and older youths. Officers and police community support officers were patrolling the streets around Parkwood Academy at Shirecliffe yesterday, with three police vehicles stationed in prominent positions. Trouble first erupted last week and since then officers have been on duty before classes, at lunchtimes, and after the final bell. Staff, including the principal Mike Westerdale, have also been on duty in streets surrounding the school. One mum said she had seen pupils fighting on the streets in incidents she believed to be racially motivated. "Our neighbourhood is being turned into a war zone," she said. "The kids go home at night but the neighbours are living with this, it's terrifying. "All the local people know it's not a good idea to go to the shops at lunchtime or after school - they are a no-go zone." A 23-year-old man who lives opposite the school on Longley Avenue West said he believed police were at the school every day in a bid to make residents feel safe. "I have seen fights and they seem to be between different racial groups. The kids seem to stick in their own racial groups and don't seem to mix much," he said. "One day I saw an officer with a big stick which he'd taken off some lad. The worst trouble seems to be between current pupils and former ones. The teachers do their best but they are obviously

---

outnumbered." A nearby shopkeeper in Teynham Road said: "All of a sudden it has flared up in the last week. I've seen all sorts of commotion and the police are trying to keep the peace as best they can." Mr Westerdale said he could only apologise to local residents and said he was personally dealing with the issues they had raised.</P>

<P>"There have been two recent assaults on Academy students by local youths," he revealed. "I am working with the police to ensure the safety of students going to and from the Academy at lunchtime and at the end of the day. "I am working closely with Sheffield Council and other community partners to help reduce the problem. "Equally, I can assure residents I will not tolerate unacceptable behaviour by students outside the Academy and a series of meetings with parents is under way." A police spokeswoman confirmed that on Tuesday at 12.20pm police were called following reports from the</P>

<P>school of a disturbance outside. "It was reported that a number of young people who were not students had attended at the school, and a confrontation took place. "Officers arrived quickly</P>

<P>at the location and dispersed the group. Inquiries are ongoing with the school and with local people, and officers are expecting to take action against identified individuals over the next few days." C h i e f</P>

<P>Insp Andy Male added: "The information we have is that this incident took place between young people from different parts of the city and is linked to where they live, rather than their ethnicity. "We are working with the school and the young people themselves, and with local residents and other agencies to ensure there is not a recurrence. "There are additional patrols around the school, consisting of PCSOs and safer neighbourhood officers. Police are maintaining a presence this week outside the premises before and after school hours to provide reassurance to school staff, pupils and parents."</P>

<P>MIKE RUSSELL News Reporter</P>

<P>'Neighbourhood turned into war zone' as trouble with pupils escalates</P>

<P>Police patrols to stop violence</P>

<P>Patrols: Police are attempting to keep the area around Parkwood Academy safe. Picture: Dean Atkins</P>

<P>Gun terror: Wedding venue.</P>

</Text>

</PageContents>

</Data>

---

## Appendix 2: Titles covered

Titles covered are listed on the [NLA Blog](#).