



## eClips Web Publisher Requirements

This document describes the inbound feed requirements.

**William Barbosa**  
**2/19/2010**

## Table of Contents

Document Control.....	4
1. Introduction .....	5
a. Intended Audience.....	5
2. Data Elements .....	5
a. Feed Type and Delivery.....	5
b. Data Requirements .....	6
a. Optional Data .....	7
3. Key Performance Indicators.....	7
4. Processes.....	8

## Document Control

Version	Date	Content/Changes	Author
Version 1.0	28/05/2010	Original Feed Requirements documents distributed to participating publisher in 2009.	Dave Goddard
Version 1.1	10/06/2010	Incorporated William Barbosa review feedback.	Stephen Vittles
Version 1.2	14/06/2010	SV updated document following review meeting.	Stephen Vittles
Version 1.3	30/06/2010	SV updated document following review meeting.	Stephen Vittles
Version 1.4	01/07/2010	Incorporated Christian's amended document.	Stephen Vittles
Version 1.5	02/07/2010	Submitted Draft for approval and sign-off.	Stephen Vittles
Version 1.6	15/07/2010	Updates and Corrections	William Barbosa

Information contained in this document is proprietary to the "Newspaper Licensing Agency Ltd".

The information contained in this document is subject to change.

## 1. Introduction

This document outlines both technical and performance requirements which the publisher must adhere to.

### a. Intended Audience

This document is technical in nature and is therefore intended for a technical audience at the Publisher providing title content for the eClips Web archive.

## 2. Data Elements

### a. Feed Type and Delivery

	Description												
Feed Type	<p>To enable NLA identification of an XML feed one of the following scenarios must be in place:</p> <p><b><u>Continuous:</u></b> A XML stream to which article objects are incrementally added. This XML feed is most commonly provided via a static HTTP URL or web service interacted with through parameters.</p> <p><b><u>Distinct XML file per Article object:</u></b> A distinct XML file per article object. Commonly, a file-naming convention such as sequential numbering is used as new article XML files are created. Typically, these types of feeds are either collected via FTP or dropped to a remote file-system.</p> <p><b><u>ZIP Files:</u></b> Zip files containing XML files. These XML files can be either multiple-article XML files or single-article XML files. Typically, these types of feeds are either collected via FTP or dropped to a remote file-system.</p>												
Content Encoding	<p>The inbound feed must be UTF-8 encoded.</p> <p>In order for valid XML, ampersand, 'less than' and 'greater than' symbols – when intended as literal characters – must use the entity escape codes listed in the table below:</p> <table border="1"> <thead> <tr> <th>Character</th> <th colspan="2">Escape Code</th> </tr> </thead> <tbody> <tr> <td>Ampersand</td> <td>&amp;</td> <td>&amp;amp;</td> </tr> <tr> <td>Greater Than</td> <td>&gt;</td> <td>&amp;gt;</td> </tr> <tr> <td>Less Than</td> <td>&lt;</td> <td>&amp;lt;</td> </tr> </tbody> </table> <p>Any other non-ASCII characters should be encoded using the <a href="#">UTF-8 codepage</a>. If this is not possible, then these characters must either be escaped using numerical entities or escaped using custom entities together with a DTD defined in the source XML.</p>	Character	Escape Code		Ampersand	&	&amp;	Greater Than	>	&gt;	Less Than	<	&lt;
Character	Escape Code												
Ampersand	&	&amp;											
Greater Than	>	&gt;											
Less Than	<	&lt;											

## b. Data Requirements

Element	Description
Article ID	The "ArticleID" is a unique key used to identify an article and its associated elements as a single unit for consumption. This can be presented in the file-name of a distinct XML file; however a literal ArticleID express in the xml file is preferable.
Version Identification	An article ID, version number, or headline that references previous versions of the same story. (This is required when using HTTP or ZIP)
Origin URI	"Origin URI" specifies the URL where the original article is found.
Published Date	Publication Date of the Article version should reflect the time the article version is released to the website. The format to be used is as follows: " <b>yyyy-mm-dd hh:mm:ss</b> ".  Later versions of an article should use the publication time of that version, and not of the original article.
Section	Information about where the article was placed.
Headline	The "Headline" element contains the full headline text.
Subheadline	The "SubHeadline" element contains the full subheadline text.
Body	The "Body" element contains the full article body text.  Body text should not be sent as escaped XML, i.e. it cannot be encapsulated within a CDATA block.  The ideal format for the body text is XHTML
Additional article Text	Article text that does not belong to the main text of the article – such as pull quotes or text associated with multimedia content – should be clearly identified with a tag or node in order to maintain the readability of the article. Text belonging to other XML fields should not be repeated in the body text.
Byline	The "Byline" element contains the Author text. It is acceptable for this field to be empty if byline text is absent from the web article.
Article Copyright	Defines the article copyright, e.g. Reuters.
Image URI / Identification	URL of the location where the image can be retrieved from. FQDN + URI is required, eg: <a href="http://www.independent.co.uk/multimedia/archive/00042/morg_42431t.jpg">http://www.independent.co.uk/multimedia/archive/00042/morg_42431t.jpg</a>  Alternatively, if the image file is supplied with the XML feed, a relative path is acceptable; for instance: "/filename.jpg" – where the image file is in the same directory as the XML file specifying it.  Images should be specified in the XML in the same order as they are displayed on the Article.
Images	All images displayed with an article must be sent in the feed. If thumbnail and teaser images must be supplied they should be clearly

	distinguished from those belonging to the article proper in the XML. Images contained in a 'Gallery' that are associated with an article but viewed in separate window are to be included in the article XML if possible, or as a separate XML file if not.
Image Caption	Defines the caption text assigned to a specific image.
Image Copyright	Defines the Copyright notice required to be displayed with images from external sources.

### a. Optional Data

Element	Description
Categories	Defines the category of the article object. Eg: News / World/ Americas.
Meta tagging	Defines the concept item of the article object. Eg: People=John McCain.
Multimedia Identification	A link to any videos associated with an article.
Multimedia transcript	A transcript of the content of any video associated with an article.
Paper Version	Reference to an appearance of the article in the print copy.
Ranking	Any ranking information applied by editorial to indicate the relative importance of an article.

## 3. Key Performance Indicators

Element	Description	Target KPI
Coverage	We require all articles published to the website to be delivered in the feed. This includes all hosted blog, third party, agency or other syndicated articles.  Any articles stored on a separate content management system or published to a separate, affiliated website (e.g. <a href="http://www.blogs.newspaper.co.uk">www.blogs.newspaper.co.uk</a> or <a href="http://www.football.newspaper.co.uk">www.football.newspaper.co.uk</a> instead of <a href="http://www.newspaper.co.uk">www.newspaper.co.uk</a> ), should be provided as a separate feed if the articles do not share an XML structure/DTD with the main article XML.	<b>KPI Target:</b> 98% <b>Definition:</b> The % of a Titles origin articles found in the eClips Web archive.
Versions	All iterations of a published article are to be made available in the feed. Any change to the article text; an addition, deletion or replacement of an image, caption, byline, copyright etc, should prompt a new article object to be sent retaining the original article ID.	All article versions
Timeliness	We require all articles and article versions to be sent to the NLA within thirty minutes of publication to the website.	<b>KPI Target:</b> 98% <b>Definition:</b> % of Archived Articles arriving within

		30 minutes of the Origin URL Article being published to the Titles website.
Consistency (Text)	Article data contained in the XML must match (be consistent with) the article at the origin URI. Where possible, inbound XML body text should match the original web copy in respect to formatting, retaining italic and bold type, paragraph indentation, line breaks, bullets etc.	<b>KPI Target:</b> 98% <b>Definition:</b> % of Archived Articles whose textual elements match those of their corresponding Origin URL Article.
Consistency (Image)	Images data contained in the XML must match (be consistent with) the images displayed with the article at the origin URI. This includes slide-shows.	<b>KPI Target:</b> 98% <b>Definition:</b> % of Archived Articles whose image objects match those of their corresponding Origin URL Article.
Field Consistency	Article XML fields must be supplied consistently, i.e. they conform to a schema.	

#### 4. Processes

Element	Description
Rights and Libel	<b>Automated restriction carried out by Publisher</b> – Publisher will apply the restriction automatically by adding to the restriction tag/field in the XML of an article. Publisher will send details of restricted articles via email for verification purposes. <b>Manual restriction carried out by NLA</b> – Publisher will send an email containing the details of an article for the NLA to restrict. <b>N.B.</b> All versions of an article will be restricted; versions received thereafter will be discarded.
Escalation process	The NLA will undertake to monitor publisher website feed(s) and will escalate directly to nominated publisher technical contacts. Should the publisher note an issue with the feed the NLA should be contacted.  The NLA ideally expects that contacts will be available 24/7/365 and can effectively deal with an escalation of a technical nature. Preferred communication method (email / telephone) to be specified by publisher.