



eClips Web API & XML Specification

A comprehensive guide to understanding and integrating with eClips Web,
the official article archive for the U.K's leading News websites.

William Barbosa

Table of Contents

Document Control.....	5
1. Introduction	6
a. Intended Audience.....	6
b. Changes.....	6
2. Overview	6
3. Licensing, Access & Authentication	7
a. Licensing & Entitlement.....	7
b. User management	7
c. Security & Authentication.....	7
a. Usage Rights & Restrictions	8
4. Payload Orchestrator	8
a. Payload API Interaction.....	8
• Authentication Method.....	9
b. Payload Interrogation Methods.....	9
c. Index-continuation method	9
• URI Template	9
• Query String Parameters	10
d. dateTime Method	10
• URI Template	10
• Query String Parameters	11
e. Payload “NewsML-G2” XML Structure.....	12
• newsMessage	12
• itemSet.....	13
• packageItem	13
• newsItem	14
• Inline XML.....	15
• xPaths to important Elements in the Payload XML.....	16
f. Payload Restriction Notification	19
5. Search.....	20

a.	Search Interaction.....	20
b.	URI Template	20
c.	Query String Parameters.....	20
d.	General treatment of Search parameters.....	21
e.	Search “Term”	21
•	Prefix terms	22
•	Noise words	22
f.	BOOLEAN Support.....	22
g.	Search Results	22
h.	Number of Results and Continuation	23
i.	Results formats	23
j.	Examples	23
•	Query strings.....	24
•	Search results as XML	24
•	Search results as HTML.....	25
•	Search results as TXT	26
6.	Redirector	26
a.	Authentication Method	26
b.	Interaction with Redirector.....	26
c.	URI Template	26
d.	Query String Parameters.....	27
7.	Article Orchestrator	27
a.	Viewing Archived Articles	27
•	As HTML.....	28
•	As PDF:.....	28
•	As Multi-article PDF	28
•	With Custom Metadata	28
•	With MMO Branding	29
b.	Authentication Method	30
c.	Interaction with Article Orchestrator (via Redirector)	30

- d. URI Template 30
- e. Query String Parameters..... 30
- Appendices..... 32
 - a. Appendix 1: What is REST? 32
 - b. Appendix 2: eClips Web “News ML-G2” Sample 32
 - c. Appendix 3: What is NITF?..... 39
 - d. Appendix 4: Website Title Codes 39
 - e. Appendix 5: Search Noise Words..... 39

Document Control

Version	Date	Content/Changes	Author
Draft 0.1	10/02/2010	Draft for internal review.	William Barbosa
Draft 0.2	18/02/2010	Integrated feedback from Business Development. Omitted sections.	William Barbosa
Draft 0.3	22/02/2010	Integrated feedback from Production. Rights and Libel Process. Restriction notification and machine processing.	William Barbosa
Draft 0.4	23/02/2010	Integrated feedback from Client Services, Business Development and Engineering. Submitted for customer review.	William Barbosa
Draft 0.5	02/03/2010	Integrated customer feedback.	Faisal Shahabuddin
Draft 0.6	02/03/2010	Integrated review feedback. Draft for BETA planning distribution. Clarified NewsML-G2 data-model Added API machine-readable response codes.	William Barbosa
Draft 0.7	23/03/2010	Several corrections. Added Custom HTTP Response Codes for Payload Orchestrator and Search API. Removed Search from Payload Orchestrator.	William Barbosa
Draft 0.8	15/09/2010	<ul style="list-style-type: none"> Relates to software build release 0.6. Updated URL's to production instances of eClips Web. Emphasised Index continuation method for Payload interrogation as the preferred access method for continual data retrieval. Licensed Entitlement Period has increased to 100 days. 	William Barbosa
Version 1.1	12/09/2011	Updated Website Title codes.	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 how media-monitoring organisation subscribers and their end-users will interact with the eClips Web platform.

a. Intended Audience

This document is technical in nature and is intended for a technical audience.

b. Changes

This document supports the BETA period and is subject to change prior to the final release of NLA eClips Web. New versions of this document will be made available on the [eClips Web blog](#).

To be notified of new updates to this document, and other eClips Web specific updates, please subscribe to the [eClips Web RSS feed](#).

2. Overview

eClips Web is an online service that provides media monitoring organisations (MMO's) with timely news content from the U.K's leading newspaper websites. This allows these MMO's to build and deliver high-quality media-monitoring solutions to their customers.

The system collects feeds generated by publishers' Content Management Systems (CMS) as it is published. Once collected the content is cleansed and normalised for archiving. This is known the cleansing and standardisation process, and this ensures that archived content is optimal for interoperability. Further, this process ensures that only the content that matters is archived, and not the "noise" (ads, user comments, navigation etc) making eClips Web the only unified original news archive of its kind in the U.K.

Further, this process is run as close to real-time as possible, although this varies from title to title. Never the less the resulting delivery of content from eClips Web is more timely than competing 'scraped' feeds.

Additionally, the NLA operate a comprehensive production monitoring and analysis process which further served to ensure the content which is collected and archived is a full representation of what publishers display on the web.

eClips Web also provides a 'permalink' facility which guarantees end-users will be able to view the content which is sent to them as a link. The eClips Web service intelligently redirects end-users (the recipients of eClips Web links) to the original article at the publisher's website, or to the eClips Web content archive, if the article has been moved or deleted from the origin website.

eClips Web uses a service oriented architecture (SOA) and is made up of four delivery components, which interact with the content archive. These are:

- **Payload Orchestrator:** A "[RESTful](#)" API, which can be programmatically interrogated and enumerates content as a "[News ML G2](#)" XML feed.
- **Search:** A "[RESTful](#)" API which allows MMO's to interrogate the eClips Web archive using Boolean search queries. Enumerates results as XML, HTML or Text.
- **Article Orchestrator:** Renders the requested article from the archive as HTML or PDF. All Article Orchestrator requests are passed through the Redirector, unless explicitly bypassed.

- **Redirector:** Directs end-user article requests to the appropriate destination, either the origin website or the archived version in the Article Orchestrator.

3. Licensing, Access & Authentication

a. Licensing & Entitlement

A “licence” defines which web-site titles a licensee has access to. Entitlement refers to a rolling period of time the licensee has access to that title for. The default entitlement period is 100 days.

b. User management

eClips Web users are managed through the existing eClips User Management Interface (<http://www.nla-eclips.com/manage>). Organisations can access this UI to manage their own and their end-user organisations.

c. Security & Authentication

eClips Web uses the following methods to authenticate users and provide access to components of the system:

Method	Characteristics
User Authentication	<ol style="list-style-type: none"> 1. Designed for browser-based interaction and authentication. 2. Users are challenged with a login dialog box and are required to provide a username and password. 3. On successful authentication, a cookie is created with a 365 day life-span and the user will not be required to re-authenticate while that cookie exists on their machine.
MMO Authentication	<ol style="list-style-type: none"> 1. Designed specifically for machine interrogation. 2. Username and Password are specified in the URL request string. 3. Only Organisation licensed as MMO’s can use services reliant on this type of authentication. 4. On successful authentication, a cookie is created with a 365 day life-span and the user/machine will not be required to re-authenticate while that cookie exists.
IP Authentication	<ol style="list-style-type: none"> 1. Access from machines originating from an allowed IP address (or range) is allowed without specifying any credentials. 2. Users can only access the service from machines originating from an allowed IP address (or range) and are still required to provide valid credentials. 3. eClips Web supports single IP addresses, a range of IP Addresses and wildcards. This method of authentication is available for both MMOs and end clients (subject to approval by the NLA). 4. On successful authentication, a cookie is created with a 365 day life-span and the user/machine will not be required to re-authenticate while that cookie exists.

All eClips Web components require authentication. Supported methods between modules vary as described in the table below.

Component	Supported Method
Payload Orchestrator	<ul style="list-style-type: none"> • MMO Authentication • IP Authentication
Search	<ul style="list-style-type: none"> • MMO Authentication • IP Authentication • User Authentication (when requesting results as HTML).
Article Orchestrator	<ul style="list-style-type: none"> • User Authentication • IP Authentication
Redirector	<ul style="list-style-type: none"> • No authentication (unless redirected to Article Orchestrator)

a. Usage Rights & Restrictions

Occasionally, publishers may restrict content for legal and other reasons. In this event, MMO's have the contractual obligation to delete any such content within 24 hours of receiving a notification.

Restriction notifications are provided in the Payload XML & this allows MMOs to automatically identify content for deletion from their systems. Further information about how these restrictions are provided in the XML can be found under the [Payload Orchestrator](#) section of this document.

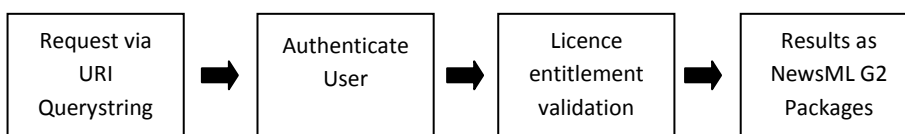
Additionally, MMO's will also be notified by e-mail that a restriction has taken place and the email instructs the MMO to remove all locally stored copies of the article (and any of its component parts) from their systems. The e-mail details the Article ID's, their Headline, version and word/character count of the Articles to be deleted.

4. Payload Orchestrator

A "RESTful" API that provides an interface to request authorised content. Content is provided as a "News ML G2" XML package allowing MMO's to integrate their applications to collect the content as required. As the Payload Orchestrator is an API designed for machine interrogation, it makes use of custom HTTP Response Codes to indicate the status of a request. These HTTP response codes and their meaning is detailed in the Query String parameter section.

a. Payload API Interaction

The diagram below represents at the basic interaction with Payload orchestrator.



The key characteristic of this interaction are that all requests for content made through the request URL. When such a request is sent to the Payload orchestrator, the API will authenticate the user (based on the supplied credentials or previously authentication where a cookie exists).

- **Authentication Method**

1. MMO Authentication.
2. Only MMO orgs can interrogate the Payload Orchestrator.

b. Payload Interrogation Methods

There are two methods to interrogate the Payload Orchestrator, each suited for specific purposes:

1. [Index-continuation Method](#): suited for continual retrieval of content as it is made available.
2. [dateTime Method](#): suited for retrieval of content published in a specified period of time.

Each method is subject to their own usage limits to ensure optimal performance of the system. These are described below in more detail.

An online reference for the parameters to interact with each of these methods can be found at the following URI:

<http://www.nla-eclipsweb.com/service/api/payload/help>

c. Index-continuation method

As previously mentioned, this method is suited for the continual retrieval of content in real-time. This method is optimal for frequent and continual polling of the eClips Web archive and offers the quickest method for retrieving manageable and predictable sets of data. Articles are ordered by the time they were received by eClips Web thereby ensuring that no articles are skipped.

- **URI Template**

The URI template describes the URI as it appears in the address space of the browser.

URI Template:	<code>http://www.nla-eclipsweb.com/service/api/payload.xml?user=[username]&index=[index Value]&rows=[Number of available articles desired]</code>
Example:	http://www.nla-eclipsweb.com/service/api/payload.xml?user=[NoSuchUser]&pwd=[NoSuchPassword]&index=[NoSuchIndexValue]&rows=100

Please note: You will have to change the User/Pwd combination in the example above to valid credentials.

- **Query String Parameters**

The table below shows the list of parameters, a description of the fields, and whether the field is optional or required.

Description	Returns Payload XML enumerated with entitled NewsML G2 packages.
Method	GET
Custom HTTP Responses	<ul style="list-style-type: none"> • 200 OK & XML (text/vnd.IPTC.NewsML). • 401 Unauthorised • 404 Not Found
Parameters	<p>Returns NewsML-G2 formatted articles.</p> <p>A list of Website Title codes for the “Title” parameter can be found in Appendix 4.</p> <p>Returns NewsML-G2 formatted articles.</p> <p>List of parameter definitions:</p> <ul style="list-style-type: none"> • Index - The payload continuation parameter. The index + 1 will be returned in the payload. (required) • Rows - The number of article objects to return. (optional) • Title - a comma separated list of title acronyms's e.g. &Title=WEBDR,WEBDM,WEBT (optional). • User - the username to authenticate. (Required for first request, thereafter can be authenticate cookie can be used). • Pwd - the password for the user. (Required for first request, thereafter can be authenticate cookie can be used).
Usage Limitations	A maximum of 100 rows can be specified. This will return 100 articles if they are available, or any number that are available.

d. dateTime Method

This method is suited for for retrieving content published in a specific period of time and is appropriate for occasional retrievals. This method in notably slower than the index-continuation method.

- **URI Template**

The URI template describes the URI as it appears in the address space of the browser.

URI Template:	http://www.nla-eclipsweb.com/service/api/payload.xml?user=[username]&pwd=[password]&start=[DD/MM/YYYY HH:MM]&end=[DD/MM/YYYY HH:MM]
Example:	http://www.nla-eclipsweb.com/service/api/payload.xml?user=[NoSuchUser]&pwd=[NoSuchPassword]&start=14/10/2010 14:00:00&end=14/10/2010 14:10:00

Please note: You will have to change the User/Pwd combination in the example above to valid credentials.

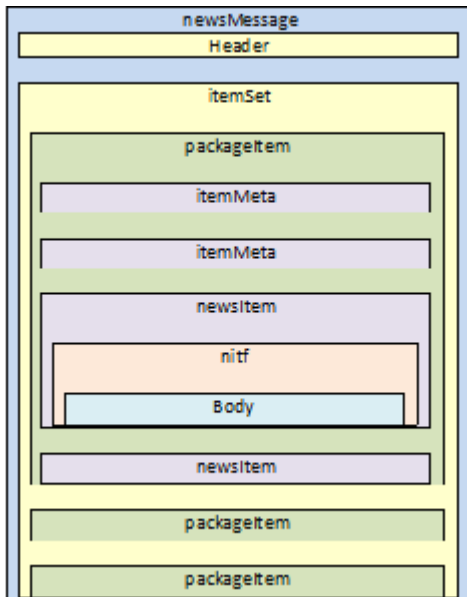
- **Query String Parameters**

The table below shows the list of parameters, a description of the fields, and whether the field is optional or required.

Description	Returns Payload XML enumerated with entitled NewsML G2 packages.
Method	GET
Custom HTTP Responses	<ul style="list-style-type: none"> • 200 OK & XML (text/vnd.IPTC.NewsML). • 204 The licensed entitlement period has been exceeded or there are no articles available in the period you have specified. Please ensure that the date of your request falls within your licensed entitlement period. • 400 Please ensure the date span does not exceed 24 hours. • 400 Please include both start and end date/times on the querystring • 401 Please make sure you enter your username (user) and password (pwd) on the QueryString.
Parameters	<p>Returns NewsML-G2 formatted articles.</p> <p>List of parameter definitions:</p> <ul style="list-style-type: none"> • Start - the start date time e.g. &start=dd/mm/yyyy 00:00:00. Date only can be used. (required) • End - the end date time e.g. &start=dd/mm/yyyy 00:00:00. Date only can be used. (required) • Title - a comma separated list of title acronyms's e.g. &Title=WEBDR,WEBDM,WEBT (optional). • User - the username to authenticate. (Required for first request, thereafter can be authenticate cookie can be used). • Pwd - the password for the user. (Required for first request, thereafter can be authenticate cookie can be used).
Usage Limitations	A maximum dateTime range of 24 hours can be specified per request for any period within the default 100 day entitlement.

e. Payload “NewsML-G2” XML Structure

The Payload Orchestrator returns a NewsML-G2 news Message (*<newsMessage>*) which contains multiple package Items or an *<itemSet>*. Each package item (*<packageItem>*) represents all objects that make up an article’s composition (in this version of eClips Web, this is limited to Textual components and Image components). The structure of the News Message for this solution is detailed in the diagram below:



- **newsMessage**

The News Message (*<newsMessage>*) defines & facilitates the exchange of any NewsItems. Each Payload request results in a newsMessage which is timestamped and can be tracked from origin.

```

- <newsMessage xsi:schemaLocation="http://iptc.org/std/nar/2006-10-01/
  http://iptc.org/std/NewsML-G2/2.2/specification/NAR_1.3-spec-
  NewsMessage-Core.xsd http://iptc.org/std/nar/2006-10-01/
  http://iptc.org/std/NewsML-G2/2.2/specification/NewsML-G2_2.2-
  spec-NewsItem-Core.xsd" xmlns="http://iptc.org/std/nar/2006-10-01/"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xmlns:nla="urn:nla-extension">
  [B] <header>
    <sent>2010-03-02T22:17:37.980951+00:00</sent>
    <sender>nla.co.uk</sender>
    <origin>NLA eClipsWeb</origin>
  </header>
  + <itemSet>
  </newsMessage>
  
```

- **itemSet**

The newsMessage is a convenient way of wrapping multiple Package Items (<packageItem>) as a single transmission within an <itemSet>. Package Items in eClips Web are all the article packages returned based on the MMO's licence entitlement (titles/days).

```
- <itemSet>
+ <packageItem>
+ <packageItem>
+ <packageItem>
</itemSet>
```

- **packageItem**

```
- <packageItem>
- <itemMeta>
  <itemClass qcode="icls:composite" />
  <link rel="irel:processedFrom"
    href="http://www.dailyrecord.co.uk/2010/03/02/kris-
    boyd-left-out-of-craig-levein-s-first-scotland-side-86908-
    22081133/" />
  <link rel="irel:associatedWith" nla:info="articleOrchestratorUrl"
    href="http://217.68.70.86/nla/redirector/article/43637.html?
    version=1" />
  <nla:articleIdentifier id="43637" version="1" />
</itemMeta>
<itemRef
  residref="urn:newsml:http://www.dailyrecord.co.uk:02/03/2010-
  18:23:00:43637:1:t"
  contentType="ApplicationException/vnd.iptc.g2.newsitem+xml" />
<itemRef
  residref="urn:newsml:http://www.dailyrecord.co.uk:02/03/2010-
  18:23:00::19465:i"
  contentType="ApplicationException/vnd.iptc.g2.newsitem+xml" />
+ <newsItem
  guid="urn:newsml:http://www.dailyrecord.co.uk:02/03/2010-
  18:23:00:43637:1:t" version="1" standard="NewsML-G2"
  standardversion="2.2" conformance="core" xsi:lang="en-us">
+ <newsItem
  guid="urn:newsml:http://www.dailyrecord.co.uk:02/03/2010-
  18:23:00::19465:i" version="1" standard="NewsML-G2"
  standardversion="2.2" conformance="core" xsi:lang="en-us">
</packageItem>
```

- **newsItem**

News Item (<newsItem>) is the way in which an individual piece of news is represented and the building block for all NewsML-G2 implementations. In eClips Web there are currently two different types of News Item that form the Package Item. News Items can be either textual, image based or composite. Each Package Item contains a single text News Item and can contain single or multiple image News Items.

Article Element	Description	xPath
Text	Identifies Textual (NITF 3.5) newsItems	newsItem[itemMeta\itemClass\@qcode=minat:text
Picture	Identifies Image (GIF/JPG/other) newsItems	newsItem[itemMeta\itemClass\@qcode=minat:picture

The News Item is the key to representing an eClips Web article object. The following table describes the XPath expressions to each logical article element in a packaged (<newsItem>):

```

- <newsItem
  guid="urn:newsml:http://www.dailyrecord.co.uk:02/03/2010-
  18:23:00:43637:1:t" version="1" standard="NewsML-G2"
  standardversion="2.2" conformance="core" xsi:lang="en-us">
  <catalogRef
    href="http://www.iptc.org/std/catalog/catalog.IPTC-G2-
    Standards_6.xml" />
  + <rightsInfo>
  + <itemMeta>
  + <contentMeta>
  + <contentSet>
  </newsItem>
  
```

Value-added Metadata

```
- <contentSet>
  <remoteContent href="/images/ImageHandler.ashx?id=19465"
    size="6843" height="100" width="100" contentType="image:JPG" />
</contentSet>
```

- **Inline XML**

The content inside <inlineXML> is valid XML and can stand alone as a valid XML document in its own namespace. In eClips Web it represents the textual components of an article.

```
- <contentSet>
  - <inlineXML contentType="application/nitf+xml" wordcount="139"
    nla:charactercount="702">
    + <nitf>
    </inlineXML>
  </contentSet>
```

In eClips Web, NITF is used as the textual content payload. NITF is used for its Text markup capability, thereby ensuring clean and simple transmission of text.

```

- <nitf>
  - <body>
    - <body.head>
      - <headline>
        <hl1>Kris Boyd left out of Craig Levein's first
          Scotland side</hl1>
      </headline>
    - <byline>
      <byttl />
    </byline>
  </body.head>
  - <body.content>
    <p>KRIS BOYD has failed to make the starting line-up
      for Scotland's friendly against Czech Republic as he
      bids to end his self-imposed international exile.</p>
    <p>The Rangers striker is set to start on the bench at
      Hampden tomorrow night after boss Craig Levein
      named the starting line-up he hopes to field.</p>
    <p>Ibrox team-mate Kenny Miller will lead the attack,
      while James McFadden and Lee Wallace have both
      been named in the team despite struggling with
      injury.</p>
    <p>If they fail to prove their fitness tomorrow
      morning, Dundee United's Paul Dixon will make his
      debut at left-back, while Graham Dorrans will
      replace McFadden.</p>
    <p>At the back, Andy Webster earns his first cap in
      four years when he partners Gary Caldwell in the
      heart of defence.</p>
    <p>Scotland team to face Czech Republic tomorrow:
      Gordon, Hutton, Caldwell, Webster, Wallace, Robson,
      Brown, D Fletcher, Thomson, McFadden, Miller.</p>
  </body.content>
</body>
</nitf>

```

- **xPaths to important Elements in the Payload XML**

For your convenience we have included the xPath's of the most required data elements of a eClips Web archived article in the following table:

Management Metadata

Article Element	Description	xPath
NLA:ArticleId	The Unique Identifier for the	/newsMessage/itemSet/packageItem/itemMeta/nla:a

entifieer	Article. The sub-node @version identifies the version of that article.	rticleIdentifier
NLA:Index	The Index Continuation value used in conjunction with the Index-continuation method .	/newsMessage/itemSet/packageItem/newsItem/itemMeta/nla:index
Title (Origin website)	The domain name of the Website which the article originated from	newsMessage/itemSet/packageItem/newsItem/itemMeta/provider
Title Acronym ID	The unique title acronym used to identify the title. Also used for title filtering methods for API interrogation.	/newsMessage/itemSet/packageItem/newsItem/itemMeta/nla:titleAcronym
Publisher Name	The name of the publisher which owns the title who distributed the article.	/newsMessage/itemSet/packageItem/newsItem/itemMeta/nla:publisher
Publishing Status	The status of the article which indicates if the article is restricted or available for use.	/newsMessage/itemSet/packageItem/newsItem/itemMeta/pubStatus

Article Data

Article Element	Description	xPath
Origin URI	The origin URI of the article	/newsMessage/itemSet/packageItem/itemMeta/link[@rel="irel:processedFrom"]
Article Orchetrator URL	The corresponding Article Orchestrator URL to display the article from the archive.	/newsMessage/itemSet/packageItem/itemMeta/link[@rel="irel:associatedWith"]
Word Count	The number of words which make up the article's textual components. Count includes	/newsMessage/itemSet/packageItem[28]/newsItem/contentSet/inlineXML/@wordcount

	count of words in Headline, Body and image captions (if any)	
Character Count	The number of characters which make up the article's textual components. Count includes count of characters in Headline, Body and image captions (if any)	/newsMessage/itemSet/packageItem[28]/newsItem/contentSet/inlineXML/@nla:charactercount
Headline	The Headline of the article	/newsMessage/itemSet/packageItem/newsItem/contentMeta/headline
Slugline	The slug line of the article, if available.	/newsMessage/itemSet/packageItem/newsItem/contentMeta/slugline
Section	The section of the website the article appeared on (if available)	/newsMessage/itemSet/packageItem/newsItem/contentMeta/nla:section
Byline (Author)	The author of the article.	/newsMessage/itemSet/packageItem/newsItem/contentSet/inlineXML/nitf/body/body.head/byline/byttl
Published Date & Time	The Date and Time that the article was published.	/newsMessage/itemSet/packageItem/newsItem/itemMeta/versionCreated
Body	The main body of the article. Refers to the textual component of the article body.	/newsMessage/itemSet/packageItem/newsItem/contentSet/inlineXML/nitf/body/body.content
Image source	The URI reference of the image which is stored and rendered from the eClips Web archive.	/newsMessage/itemSet/packageItem/newsItem/contentSet/remoteContent
Image caption	The textual caption related to the Image object (if present)	/newsMessage/itemSet/packageItem/newsItem/contentMeta/description[@role="drol:caption"]
Image credit	The copyright or attribution textual element related to	/newsMessage/itemSet/packageItem/newsItem/contentSet/inlineXML/nitf/body/body.head/byline/byttl

	the image.	entMeta/creditline
--	------------	--------------------

f. Payload Restriction Notification

NewsML-G2 provides information on the publishing status of an article. This is referred to as pubStatus.

By default, and indicating that an article object is available for processing and/or storage the pubStatus of an Article is: pubStatus = Usable

```

- <itemMeta>
  <itemClass qcode="ninat:text" />
  <provider literal="http://www.dailymail.co.uk" />
  <versionCreated>2010-03-02T07:01:00</versionCreated>
  <pubStatus qcode="stat:usable" />
  <nla:publisher literal="Associated Newspapers Ltd" />
  <nla:restrictionStatus literal="" />
</itemMeta>

```

When an Article is restricted by the Publisher the pubStatus of the restricted article changes to: pubStatus = Withdrawn.

This instructs receiving systems to remove all versions of the article from all locations.

The Payload Orchestrator “start” and “end” parameters also request all restrictions applied during the requested period and these will be served along with articles published and “usable” during that period.

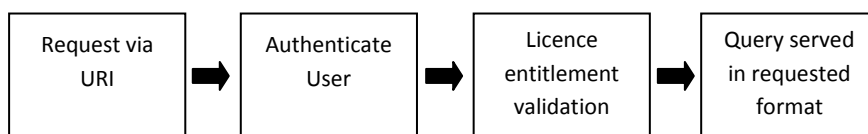
newsItem Status	Description	xPath
Publishing Status indicating usable	Refers to the Usable restriction state of the Article.	newsItem\itemMeta\pubStatus\@qcode=stat:usable
Publishing Status indicating restriction	Refers to the Restricted state of the Article and instructs the machine to delete all copies.	newsItem\itemMeta\pubStatus\@qcode=stat:cancelled

5. Search

Search allows MMO's to search against the entitled content for textual matches. The Search API returns a list of articles within MMO entitlement that matches the search term(s) specified. Text search is performed on the Headline, Byline, and Body fields of all the entitled article versions.

When requesting search result as XML, the API assumes this is for machine interrogation and therefore responds using custom HTTP Response Codes which are detailed in the Querystring Parameters. However, if requesting search results as either HTML or TXT, the API responds using standard HTTP Responses and human-readable HTML messages.

a. Search Interaction



b. URI Template

The URI template describes the URI as it appears in the address space of the browser.

URI Template:	<code>http://www.nla-eclipsweb.com/service/search/articles.xml?term=[BOOLEAN]&user=[username]&pwd=[password]</code>
Example:	http://www.nla-eclipsweb.com/service/search/articles.xml?term=Apple NOT iPod&user=NoSuchUser&pwd=NoSuchPassword

Please note: You will have to change the User/Pwd combination in the example above to valid credentials.

c. Query String Parameters

The table below shows the list of parameters, a description of the fields, and whether the field is optional or required.

Description	Returns a list of articles where the Body, Headline, or Byline match the search term, and match the parameters as described below.
URI Template	<code>http://www.nla-eclipsweb.com/service/Search/Articles.xml?term=[search]&Title=[title]&start=[start]&end=[end]&from=[fromRowNumber]&rows=[rowCount]&user=[user]&pwd=[pwd]</code>
Method	GET
Custom HTTP Responses	<ul style="list-style-type: none"> • 200 OK • 400 Bad Request, Either the start date is after the end date, the date was not in the correct format, or the search term is empty. • 204 No Content, No results returned. • 401 Unauthorised, The user credentials are missing or invalid. • 500 Internal Server Error, Unhandled exception.

Parameters	A list of Website Title codes for the “Title” parameter can be found in Appendix 4 .		
	<u>List of parameter definitions:</u>		
	Parameter	Description	Required?
	Term	The search term (must not be empty), Boolean operators AND, OR, NOT and phrase grouping e.g. &Term="Arsenal" NOT "Football".	Required
	Title	A comma separated list of title codes e.g. &Title=WEBDM,WEBDT, WEBDMI	Optional
	Start	The start date time e.g. &start=dd/mm/yyyy hh:mm:ss. Date only can be used.	Optional
	End	The end date time e.g. &start=dd/mm/yyyy hh:mm:ss. Date only can be used.	Optional
	From	The continuation index of the list of articles returned.	Optional
	Rows	The number of results per page returned	Optional
	User	The username to authenticate. Required for every request.	Required
Pwd	The password for the user. Required for every request.	Required	

d. General treatment of Search parameters

- If the “term” parameter is empty or not specified an error message is presented to the user.
- If the “start” parameter is not specified the start date will be 28 days (the default entitlement period) from the current date and time.
- If the “end” parameter is not specified the end date will be the current date and time.
- The username and password is authenticated every time a search request is performed and if authentication fails or authentication details are not provided an error is logged and a response is returned to the user.
- The username and password is saved as cookies by the redirector when supplied in the query string. The username and password cookies are then retrieved by the search service and used in the authentication process. When a user is authenticated information is returned to identify the user and to determine which articles are available due to restriction or licensing (see Entitlement and Restrictions/Embargo).

e. Search “Term”

The search uses an index to generate results based on the term supplied in the query string. Some general best-practice should be observed when performing searches.

Boolean Operators must be in upper case and these include AND, OR, NOT, AND NOT (same as NOT).

Phrase grouping is provided for by the use of double quotes around each phrase. The search is robust enough to deal with missing quotes and will insert these before and after Boolean operators and beginning and end of the text where appropriate.

- **Prefix terms**

This relates to a word or phrase where the words begin with specified text (not case sensitive). This is performed by using the “*” at the end of a term. E.g. Term=”tech*” would return results that began with tech e.g. “Technology” and “Technical” etc. The asterisk cannot be placed at the beginning or in the middle of a term but must always be at the end. Multiple wildcards can be used in multiple phrases e.g. Term=tech* OR info*.

- **Noise words**

Noise words are words that the Search API ignores as search terms e.g. while, which etc. A list of all noise words are found in [Appendix 4](#).

f. BOOLEAN Support

Search performs text matching against the Headline, Byline and Body elements of articles and supports basic BOOLEAN logic:

- Chaining operations using AND/OR/NOT operators.

```
http://www.nla-eclipsweb.com/service/search/articles.html?term=apple AND iPad AND NOT iPod
```

- Phrase Grouping is provided through the use of quotation marks “” around each phrase.

```
http://www.nla-eclipsweb.com/service/search/articles.html?term="apple launch iPad" AND "iPod reviews"
```

g. Search Results

The Search API allows MMO’s to integrate search functions directly against the eClips Web article archive. This service returns a simplified set of results allowing lightweight search interoperability. These are:

- **Article ID:** The unique eClips Web ID of the matching article.
- **Headline:** The Headline of the matching article.
- **Publisher:** The Publisher of the Website title.
- **Title:** The Website title of the matching article.
- **Publication Date:** The original Publication date of the matching article.
- **Byline:** The Author of the matching article.
- **Version:** The Version of the matching article.

- **Word Count:** The number of words in the Body of the Article.
- **URL:** The Article Orchestrator URI for the matching article.

```
http://www.nla-eclipsweb.com/service/search/articles.xml?term=[BOOLEAN]
```

h. Number of Results and Continuation

By default the search results returned per page is 25. The number of results can be altered from the default by specifying the rows parameter.

```
http://www.nla-eclipsweb.com/service/search/articles.html?term=[BOOLEAN]&rows=10
```

If your search results in more than either the default number or specified number of results displayed, the from parameter is used as a continuation parameter.

```
http://www.nla-eclipsweb.com/service/search/articles.html?term=[BOOLEAN]&rows=10&from=11
```

i. Results formats

The simple Search API also provides results in different formats, these are:

- **HTML:** Search results as HTML

```
http://www.nla-eclipsweb.com/service/search/articles.html?term=[BOOLEAN]
```

- **XML:** Search results as XML

```
http://www.nla-eclipsweb.com/service/search/articles.xml?term=[BOOLEAN]
```

- **TXT:** Search results as ASCII

```
http://www.nla-eclipsweb.com/service/search/articles.txt?term=[BOOLEAN]
```

j. Examples

- Query strings

Operator	String Example	eCW Search Query
AND	ps3 AND bug	http://www.nla-eclipsweb.com/service/search/articles.xml?term=ps3 AND bug
NOT	ps3 NOT bug	http://www.nla-eclipsweb.com/service/search/articles.xml?term=ps3 NOT bug
OR	ps3 AND (playstation3 OR playstation 3)	http://www.nla-eclipsweb.com/service/search/articles.xml?term=ps3 AND (playstation3 OR playstation 3)
AND NOT	ps3 AND bug AND NOT sony	http://www.nla-eclipsweb.com/service/search/articles.xml?term=ps3 AND bug AND NOT sony
PHRASE GROUPING	"Leap year bug hits games consoles" OR "Don't use your PS3, warns Sony"	<a \"don't="" bug="" consoles\"="" games="" hits="" href="http://www.nla-eclipsweb.com/service/search/articles.xml?term=\" leap="" or="" ps3,="" sony\""="" use="" warns="" year="" your="">http://www.nla-eclipsweb.com/service/search/articles.xml?term=\"Leap year bug hits games consoles\" OR \"Don't use your PS3, warns Sony\"
WILDCARD	labour*	http://www.nla-eclipsweb.com/service/search/articles.xml?term=labour*

- Search results as XML

```

<Articles start="02/02/2010 11:29:25" end="02/03/2010 11:29:25" term="ps3 AND bug"
results="12" xmlns="URN:Auros-xArticles-Schema"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
- <Article Id="43061">
  <Headline>Sony fix PlayStation 3 technical problem</Headline>
  <Publisher>Trinity Mirror plc</Publisher>
  <Title>Mirror.co.uk</Title>
  <PublishedDate>02 March 2010</PublishedDate>
  <Byline />
  <Version>1</Version>
  <WordCount>177</WordCount>
  <ArticleUri>http://217.68.70.86/nla/redirector/article/43061.html</ArticleUri>
</Article>
+ <Article Id="43030">
+ <Article Id="42930">
+ <Article Id="41908">
+ <Article Id="42926">
+ <Article Id="43057">
+ <Article Id="42077">
+ <Article Id="42608">
+ <Article Id="41983">
+ <Article Id="41866">
+ <Article Id="41745">
+ <Article Id="41627">
</Articles>

```

- [Search results as HTML](#)

eClips Web Search Results

Total search results: 13

Article Id: 43144
Headline: Electrical giant fixes PS3 glitch
Publisher: DC Thomson & Co Ltd
Title: Eveningexpress.co.uk
Published Date: 02 March 2010
Byline:
Version: 1
Word Count: 106
Url: <http://217.68.70.86/nla/redirector/article/43144.html>

Article Id: 43061
Headline: Sony fix PlayStation 3 technical problem
Publisher: Trinity Mirror plc
Title: Mirror.co.uk
Published Date: 02 March 2010
Byline:
Version: 1
Word Count: 177
Url: <http://217.68.70.86/nla/redirector/article/43061.html>

Article Id: 43030
Headline: Sony PS3 problems fixed
Publisher: Telegraph Media Group Limited
Title: Telegraph.co.uk
Published Date: 02 March 2010
Byline:
Version: 1
Word Count: 262
Url: <http://217.68.70.86/nla/redirector/article/43030.html>

- **Search results as TXT**

```

NLA Search Result
Results returned: 19
text=ps3 AMD bug

Article Id: 43144
Headline: Electrical giant fixes PS3 glitch
Publisher: DC Thomson & Co Ltd
Title: Eveningexpress.co.uk
Publisher Date:
Byline:
Version: 1
Word Count: 106
Url: http://217.68.70.86/nla/redirector/article/43144.html

Article Id: 43061
Headline: Sony fix PlayStation 3 technical problem
Publisher: Trinity Mirror plc
Title: Mirror.co.uk
Publisher Date:
Byline:
Version: 1
Word Count: 177
Url: http://217.68.70.86/nla/redirector/article/43061.html

Article Id: 43030
Headline: Sony PS3 problems fixed
Publisher: Telegraph Media Group Limited
Title: Telegraph.co.uk
Publisher Date:
Byline:
Version: 1
Word Count: 162
Url: http://217.68.70.86/nla/redirector/article/43030.html
    
```

6. Redirector

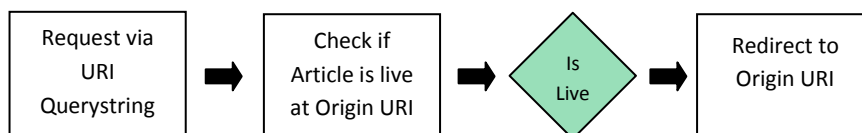
The redirector acts as a permalink proxy for Websites available on eClips Web. Every time a user requests an eClips Web article through the redirector, the redirector validates that the content hosted at the publisher’s Original URI is the same as the article expected. If the article at the Origin URI is found to be the same, the user is directed to the Origin URI, if not the user is directed to the Article Orchestrator.

a. Authentication Method

No Authentication required if article in Origin URI is found to be the same as the archived article.

Users directed to the Article Orchestrator using the Redirect method, will be required to authenticate.

b. Interaction with Redirector



c. URI Template

The URI template describes the URI as it appears in the address space of the browser.

http://www.nla-eclipsweb.com/service/redirector/article/{articleId}.html

d. Query String Parameters

The table below shows the list of parameters, a description of the fields, and whether the field is optional or required.

Description	Redirects the user to the article, either the Original Article at the publisher website if available in the public domain (including content hosted behind a pay-wall)		
Method	GET		
HTTP Responses	<ul style="list-style-type: none"> • 200 OK • 400 Bad Request • 401 Unauthorised • 404 Article Not Found 		
Parameters	List of parameter definitions:		
	Parameter	Description	Required?
	ArticleId	eClips Web ID of article to be viewed.	Required

7. Article Orchestrator

The Article Orchestrator renders an easy to read HTML representation of the archived Article, or set of articles (depending on the file-type extension).

The Article Orchestrator is accessed when:

- The expected Article is no longer available at the Origin URI or is offline.
- Directly by requesting an article PDF or News ML newsMessage by file-type extension.
- Directly by requesting a specific Version of an article with any extension

Extensions allow the user to view the archived article in a number of ways. These are

- HTML – XHTML representation with Navigation bar. Is optimised for Web browsing and is constrained to a single article per request.
- PDF – PDF of the XHTML representation above optimized for A4 printing. The MMO has the option of “Branding” this output with a specific layout & logo as explained in the branding section below

a. Viewing Archived Articles

Archived Articles can be presented in several format and by specifying the desired file-type extension in the URI request string, the Article Orchestrator will render the article in that format. These are:

- **As HTML**

xHTML representation with contextual navigation bar. Is optimised for browser viewing and is constrained to a single article per request. Unless the version parameter is specified, will run an Origin URI validation.

```
http://www.nla-eclipsweb.com/service/redirector/article/[articleID].html
```

- **As PDF:**

PDF of the XHTML representation above optimized for A4 paper printing. Acts as an Origin URI validation bypass. If no version parameter is specified, the latest archived article will be displayed. This format also supports multi-object calls that is detailed separately.

```
http://www.nla-eclipsweb.com/service/redirector/article/[articleID].pdf
```

- **As Multi-article PDF**

Requests a PDF rendering of many articles from different website titles together as a single file. Both the ArticleID and Version of the required articles must be specified and these parameters are separated by a colon (:). Each Article is the separated by a comma (,).

```
http://www.nla-eclipsweb.com/service/redirector/article/articles.pdf?articles=[articleID]:[version],[articleID]:[version],[articleID]:[version]
```

- **With Custom Metadata**

Provision is made for MMO's to pass text data to the Article Orchestrator to be displayed in the Metadata container of the archives article.



Consideration must be made with the number of metadata items you want to display, as this will affect the layout of the archived article.

This data is provided via the request string and each data entry is separated by semi-colon.

```
http://www.nla-eclipsweb.com/service/redirector/article/[articleID].html?version=[version number]& meta=[MetaData 1];[MetaData 2]; [MetaData 3];[MetaData 4]; [MetaData 5]...
```

Using a colon (:) as a suffix to a meta-term will instruct the Article Orchestrator to display the term in **BOLD**. Eg:

```
http://www.nla-eclipsweb.com/service/redirector/article/[articleID].html?version=[version number]& meta=[Meta1: Hello World 1];[Meta2: Hello World 2];[Meta3: Hello World 3]...
```

- **With MMO Branding**

By specifying the OrgID value, the Article Orchestrator can be made to render an organisation - specific Logo in the upper right hand corner of the archived article display. Only one Unique Logo per Organisation is permitted. Branding can be applied to both MMO organisations and their Customer Organisations.

If no branding for the OrgID specified has been configured, the Article Orchestrator will default to the eClips Web Logo.

To have branding associated with a particular organisation, please contact your eClips Web customer account manager or NLA client services.

```
http://www.nla-eclipsweb.com/service/redirector/article/[articleID].html?version=[version number]&orgid=[orgid]
```

eClips Web Branding (Default)





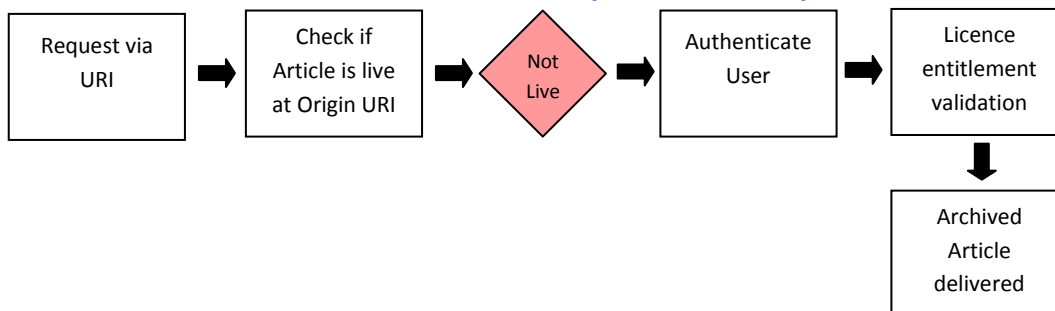
b. Authentication Method

User Authentication

If no credentials are provided in the request string, then the User will be challenged for their credentials.

On successful authentication, cookie is stored on user machine for 365 days, during which time the user does not have to provide security credentials.

c. Interaction with Article Orchestrator (via Redirector)



d. URI Template

The URI template describes the URI as it appears in the address space of the browser.

```
http://www.nla-eclipsweb.com/service/redirector/article/{articleId}.{extension}?version={version}&orgid={orgid}&user={user}&pwd={pwd}
```

e. Query String Parameters

The table below shows the list of parameters, a description of the fields, and whether the field is optional or required.

Description	Returns a list of articles where the Body, Headline, or Byline match the search term, and match the parameters as described below.
Method	GET
HTTP	<ul style="list-style-type: none"> 200 OK

Responses	<ul style="list-style-type: none"> • 400 Bad Request • 401 Unauthorised • 404 Article Not Found 												
Parameters	<p>List of parameter definitions:</p> <table border="1"> <thead> <tr> <th data-bbox="459 501 603 533">Parameter</th> <th data-bbox="603 501 1217 533">Description</th> <th data-bbox="1217 501 1474 533">Required?</th> </tr> </thead> <tbody> <tr> <td data-bbox="459 533 603 564">ArticleId</td> <td data-bbox="603 533 1217 564">eClips Web ID of article to be viewed.</td> <td data-bbox="1217 533 1474 564">Required</td> </tr> <tr> <td data-bbox="459 564 603 730">Version</td> <td data-bbox="603 564 1217 730">The version of the article to be viewed. Will default to latest version if not specified. Also used as a redirection bypass parameter. If not specified the Redirector will validate the Origin URI against the archived article and the user may be redirected.</td> <td data-bbox="1217 564 1474 730">Optional</td> </tr> <tr> <td data-bbox="459 730 603 831">OrgID</td> <td data-bbox="603 730 1217 831">The eClips OrgId defines Organsation-specific custom branding logo to use. If no OrgID is specified defaults to eClips Web branding.</td> <td data-bbox="1217 730 1474 831">No.</td> </tr> </tbody> </table>	Parameter	Description	Required?	ArticleId	eClips Web ID of article to be viewed.	Required	Version	The version of the article to be viewed. Will default to latest version if not specified. Also used as a redirection bypass parameter. If not specified the Redirector will validate the Origin URI against the archived article and the user may be redirected.	Optional	OrgID	The eClips OrgId defines Organsation-specific custom branding logo to use. If no OrgID is specified defaults to eClips Web branding.	No.
Parameter	Description	Required?											
ArticleId	eClips Web ID of article to be viewed.	Required											
Version	The version of the article to be viewed. Will default to latest version if not specified. Also used as a redirection bypass parameter. If not specified the Redirector will validate the Origin URI against the archived article and the user may be redirected.	Optional											
OrgID	The eClips OrgId defines Organsation-specific custom branding logo to use. If no OrgID is specified defaults to eClips Web branding.	No.											

Appendices

a. Appendix 1: What is REST?

REST is a term used to describe an architecture style for networked systems. REST is an acronym which stands for “Representational State Transfer”. To summarise rest with an example, The Web is a REST system. By this we mean that REST operates on the same principles and has the same characteristics as the web. Each resource available in the eClips Web system is defined and accessible through a URI. By definition REST is stateless which means all information is processed a single unique request.

b. Appendix 2: eClips Web “News ML-G2” Sample

NewsML-G2 is the XML multimedia news exchange format standard of the [IPTC](http://www.iptc.org), the International Press Telecommunications Council.

As part of the IPTC G2-Standards family, NewsML-G2 acts as an envelope for one or more news items (such as a text article, a photo, or a video clip) or a structured package of links to news items, and contains metadata to describe the relationships between the items. NewsML-G2 is a business-to-business standard that is intended to help news agencies create complex packages of multimedia news into a single cohesive bundle.

```

newsMessage xsi:schemaLocation="http://iptc.org/std/nar/2006-10-01/
http://iptc.org/std/NewsML-G2/2.2/specification/NAR_1.3-spec-
NewsMessage-Core.xsd http://iptc.org/std/nar/2006-10-01/
http://iptc.org/std/NewsML-G2/2.2/specification/NewsML-G2_2.2-spec-
NewsItem-Core.xsd" xmlns="http://iptc.org/std/nar/2006-10-01/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns:nla="urn:nla-extension">
header>
sent>2010-03-24T10:33:17.1640531+00:00</sent>
sender>nla.co.uk</sender>
origin>NLA eClipsWeb</origin>
</header>
itemSet>
packageItem>
itemMeta>
itemClass qcode="icls:composite" />
link rel="irel:processedFrom" href="http://www.dailymail.co.uk/news/article-
1260222/Police-quiz-airport-X-ray-security-worker-used-naked-body-scanner-
leer-colleague.html" />
link rel="irel:associatedWith" nla:info="articleOrchestratorUrl"
href="http://www.nla-

```

```

eclipsweb.com/service/redirector/article/86701.html?version=15" />
nla:articleIdentifier id="86701" version="15" />
</itemMeta>
itemRef residref="urn:newsml:http://www.dailymail.co.uk:24/03/2010-
10:20:00:86701:15:t"
contenttype="ApplicationException/vnd.iptc.g2.newsitem+xml" />
itemRef residref="urn:newsml:http://www.dailymail.co.uk:24/03/2010-
10:20:00:None:122766:i"
contenttype="ApplicationException/vnd.iptc.g2.newsitem+xml" />
itemRef residref="urn:newsml:http://www.dailymail.co.uk:24/03/2010-
10:20:00:AFP/Getty-Images:122767:i"
contenttype="ApplicationException/vnd.iptc.g2.newsitem+xml" />
itemRef residref="urn:newsml:http://www.dailymail.co.uk:24/03/2010-
10:20:00:None:122768:i"
contenttype="ApplicationException/vnd.iptc.g2.newsitem+xml" />
itemRef residref="urn:newsml:http://www.dailymail.co.uk:24/03/2010-
10:20:00:None:122769:i"
contenttype="ApplicationException/vnd.iptc.g2.newsitem+xml" />
itemRef residref="urn:newsml:http://www.dailymail.co.uk:24/03/2010-
10:20:00:None:122770:i"
contenttype="ApplicationException/vnd.iptc.g2.newsitem+xml" />
itemRef residref="urn:newsml:http://www.dailymail.co.uk:24/03/2010-
10:20:00:AFP/Getty-Images:122771:i"
contenttype="ApplicationException/vnd.iptc.g2.newsitem+xml" />
newsItem guid="urn:newsml:http://www.dailymail.co.uk:24/03/2010-
10:20:00:86701:15:t" version="15" standard="NewsML-G2" standardversion="2.2"
conformance="core" xsi:lang="en-us">
catalogRef href="http://www.iptc.org/std/catalog/catalog.IPTC-G2-
Standards_6.xml" />
rightsInfo>
copyrightHolder literal="Dailymail.co.uk" />
</rightsInfo>
itemMeta>
itemClass qcode="ninat:text" />
provider literal="http://www.dailymail.co.uk" />
versionCreated>2010-03-24T10:20:00</versionCreated>
pubStatus qcode="stat:usable" />
nla:publisher literal="Associated Newspapers Ltd" />
nla:restrictionStatus literal="" />
</itemMeta>
contentMeta>
creator literal="Sun" />
language tag="en-GB" />
slugline>Scott Warren</slugline>
headline>Police quiz airport X-ray security worker 'who used naked body scanner
to leer at colleague'</headline>
nla:AVE>Currently Not Implemented</nla:AVE>
nla:ABCe>
CertificateUrl

```

```

xmlns="">http://www.abc.org.uk/Certificates/16683040.pdf</CertificateUrl>
ReturnNumber xmlns="">16683040</ReturnNumber>
Primary xmlns="">2156077</Primary>
Secondary xmlns="" />
</nla:ABCe>
</contentMeta>
contentSet>
inlineXML contentType="application/nitf+xml" wordcount="737"
  nla:charactercount="3814">
nitf>
body>
body.head>
hedline>
h1>Police quiz airport X-ray security worker 'who used naked body scanner to
leer at colleague'</h1>
</hedline>
byline>
byttl>Scott Warren</byttl>
</byline>
</body.head>
body.content>
p>Police quiz airport X-ray security worker 'who used naked body scanner to
leer at colleague' An airport security guard who made lewd comments about a
female colleague's breasts after taking a 'naked' body scan of her is facing the
sack.John Laker, 25, took the image at Heathrow airport after Jo Margetson
walked into the machine by accident. He then said: ' I love those gigantic
****'. 'Miss Margetson, 29, found the remark offensive and immediately
complained to managers at BAA, which runs Heathrow, and the police.'Ican't
bear to think about the body scanner thing,' she said. 'I'm totally traumatised
by it.'I've spoken to the police about it. I'm in too much of a state to go to
work.'Laker - who is the first airport worker to be investigated for misusing the
device - was given a police warning for harassment.The body scanners have
been introduced at Manchester and Heathrow after last year's unsuccessful
ChristmasDay bombing on a flight to Detroit. They are expected to be rolled out
atBirmingham and Gatwick later this year.The machines produce an image of
the subject's body that clearly shows the outline of the genitals and other body
parts.A police spokesman said: 'Police received an allegation of harassment
regarding an incident at Heathrow Terminal 5 on March 10. 'A first instance
harassment warning has been issued to a 25-year-old man.'A spokesman for
BAA: 'We treat any allegations of inappropriate behaviour or misuse of security
equipment very seriously and these claims are investigated thoroughly.'If these
claims are found to be substantiated, we will take appropriate action.' The
incident has prompted renewed calls for the controversial apparatus to be
scrapped.Alex Deane, from civil liberties group Big Brother Watch, said:&nbsp;
'When privacy-invading machines are installed at our airports, abuses like this
are inevitable.'We are disturbed to hear of this case, which will lead to greater
calls for these scanners to be banned.'For every official caught ogling like this,
there are plenty more eyeing-up law-abiding travellers. These expensive
machines are totally disproportionate.'However, transport&nbsp; secretary
    
```

Lord Adonis resolutely backed the machines. 'A pat-down search is intrusive, but people accept that. A body scanner is in the same category,' he said. 'Images are deleted and staff are properly trained and supervised.' He was echoed by Commons Home Affairs Committee chair Keith Vaz, who said 'Body scanners are a vital resource in fighting terrorism'. He added anyone caught abusing their access to the images should receive 'sufficient punishment'. Controversy has dogged the devices since a trial began at Manchester airport, which led to the first machine entering service on February 5. Less than a month later, the first people to refuse a scan on religious and health grounds - two women travelling to Pakistan - were told they could not fly. That followed a directive from Islamic scholars in the U.S. that Muslim travellers should not pass through the scanners because they violate religious rules on nudity. The question of privacy was raised in a report on airport security by the House of Commons Home Affairs Committee today. The committee said: 'Having witnessed these full-body scanners working at first hand, we are confident that the privacy concerns that have been expressed in relation to these devices are overstated and that full body scanners are no more an invasion of privacy than manual 'pat-downs' or searches of bags.' Air passengers already tolerate a large invasion of their privacy and we do not feel that full body scanners add greatly to this situation. Privacy concerns should not prevent the deployment of scanners. 'One of the bodies that has questioned the legality of scanners is the Equality and Human Rights Commission. Responding to the Home Affairs Committee report, the commission said: 'We believe there is a risk that the way body scanning was introduced in UK airports breaches discrimination law, as well as breaching passengers' human right to privacy. 'We are concerned that a lack of training for staff who undertake these scans, as well as a lack of safeguards, mean that authorities will be unable to check the system is being operated lawfully, with fairness and without discriminating against people just because of their race, religion, gender, age, sexual orientation or disability. 'Of course, breaching people's right to privacy is justifiable when you are protecting passengers' safety and security, but it has to be within the law. 'We believe the Government has failed to demonstrate that this policy is within the law, which is why we wrote to (Home Secretary) Alan Johnson earlier this year with our concerns.'

</p>

</body.content>

</body>

</nitf>

</inlineXML>

</contentSet>

</newsItem>

newsItem guid="urn:newsml:http://www.dailymail.co.uk:24/03/2010-

10:20:00:None:122766:i" version="1" standard="NewsML-G2"

standardversion="2.2" conformance="core" xsi:lang="en-us">

catalogRef href="http://www.iptc.org/std/catalog/catalog.IPTC-G2-

Standards_6.xml" />

rightsInfo>

copyrightHolder literal="Dailymail.co.uk" />

</rightsInfo>

temMeta>

```

itemClass qcode="ninat:picture" />
provider literal="http://www.dailymail.co.uk" />
versionCreated>2010-03-24T10:20:00</versionCreated>
pubStatus qcode="stat:usable" />
</itemMeta>
contentMeta>
language tag="en-GB" />
creditline>None</creditline>
description role="drol:caption">'Traumatised': Jo Margetson complained after a
colleague made remarks about her body</description>
</contentMeta>
contentSet>
remoteContent href="/images/ImageHandler.ashx?id=122766" size="16514"
height="301" width="233" contenttype="image:JPG" />
</contentSet>
</newsItem>
newsItem guid="urn:newsml:http://www.dailymail.co.uk:24/03/2010-
10:20:00:AFP/Getty-Images:122767:i" version="1" standard="NewsML-G2"
standardversion="2.2" conformance="core" xsi:lang="en-us">
catalogRef href="http://www.iptc.org/std/catalog/catalog.IPTC-G2-
Standards_6.xml" />
rightsInfo>
copyrightHolder literal="Dailymail.co.uk" />
</rightsInfo>
itemMeta>
itemClass qcode="ninat:picture" />
provider literal="http://www.dailymail.co.uk" />
versionCreated>2010-03-24T10:20:00</versionCreated>
pubStatus qcode="stat:usable" />
</itemMeta>
contentMeta>
language tag="en-GB" />
creditline>AFP/Getty Images</creditline>
description role="drol:caption">Graphic images: A member of staff at Manchester
Airport demonstrates the scanner</description>
</contentMeta>
contentSet>
remoteContent href="/images/ImageHandler.ashx?id=122767" size="20798"
height="423" width="233" contenttype="image:JPG" />
</contentSet>
</newsItem>
newsItem guid="urn:newsml:http://www.dailymail.co.uk:24/03/2010-
10:20:00:None:122768:i" version="1" standard="NewsML-G2"
standardversion="2.2" conformance="core" xsi:lang="en-us">
catalogRef href="http://www.iptc.org/std/catalog/catalog.IPTC-G2-
Standards_6.xml" />
rightsInfo>
copyrightHolder literal="Dailymail.co.uk" />
    
```

```

</rightsInfo>
itemMeta>
itemClass qcode="ninat:picture" />
provider literal="http://www.dailymail.co.uk" />
versionCreated>2010-03-24T10:20:00</versionCreated>
pubStatus qcode="stat:usable" />
</itemMeta>
contentMeta>
language tag="en-GB" />
creditline>None</creditline>
description role="drol:caption">'Traumatised': Airport worker Jo Margetson was
allegedly leered at by a colleague</description>
</contentMeta>
contentSet>
remoteContent href="/images/ImageHandler.ashx?id=122768" size="6492"
height="115" width="154" contenttype="image:JPG" />
</contentSet>
</newsItem>
newsItem guid="urn:newsml:http://www.dailymail.co.uk:24/03/2010-
10:20:00:None:122769:i" version="1" standard="NewsML-G2"
standardversion="2.2" conformance="core" xsi:lang="en-us">
catalogRef href="http://www.iptc.org/std/catalog/catalog.IPTC-G2-
Standards_6.xml" />
rightsInfo>
copyrightHolder literal="Dailymail.co.uk" />
</rightsInfo>
itemMeta>
itemClass qcode="ninat:picture" />
provider literal="http://www.dailymail.co.uk" />
versionCreated>2010-03-24T10:20:00</versionCreated>
pubStatus qcode="stat:usable" />
</itemMeta>
contentMeta>
language tag="en-GB" />
creditline>None</creditline>
description role="drol:caption">'Traumatised': Airport worker Jo Margetson was
allegedly leered at by a colleague</description>
</contentMeta>
contentSet>
remoteContent href="/images/ImageHandler.ashx?id=122769" size="11479"
height="185" width="308" contenttype="image:JPG" />
</contentSet>
</newsItem>
newsItem guid="urn:newsml:http://www.dailymail.co.uk:24/03/2010-
10:20:00:None:122770:i" version="1" standard="NewsML-G2"
standardversion="2.2" conformance="core" xsi:lang="en-us">
catalogRef href="http://www.iptc.org/std/catalog/catalog.IPTC-G2-
Standards_6.xml" />

```

```

rightsInfo>
copyrightHolder literal="Dailymail.co.uk" />
</rightsInfo>
itemMeta>
itemClass qcode="ninat:picture" />
provider literal="http://www.dailymail.co.uk" />
versionCreated>2010-03-24T10:20:00</versionCreated>
pubStatus qcode="stat:usable" />
</itemMeta>
contentMeta>
language tag="en-GB" />
creditline>None</creditline>
description role="drol:caption">'Traumatised': Airport worker Jo Margetson was
allegedly leered at by a colleague</description>
</contentMeta>
contentSet>
remoteContent href="/images/ImageHandler.ashx?id=122770" size="4104"
height="84" width="87" contenttype="image:JPG" />
</contentSet>
</newsItem>
newsItem guid="urn:newsml:http://www.dailymail.co.uk:24/03/2010-
10:20:00:AFP/Getty-Images:122771:i" version="1" standard="NewsML-G2"
standardversion="2.2" conformance="core" xsi:lang="en-us">
catalogRef href="http://www.iptc.org/std/catalog/catalog.IPTC-G2-
Standards_6.xml" />
rightsInfo>
copyrightHolder literal="Dailymail.co.uk" />
</rightsInfo>
itemMeta>
itemClass qcode="ninat:picture" />
provider literal="http://www.dailymail.co.uk" />
versionCreated>2010-03-24T10:20:00</versionCreated>
pubStatus qcode="stat:usable" />
</itemMeta>
contentMeta>
language tag="en-GB" />
creditline>AFP/Getty Images</creditline>
description role="drol:caption">Invasive: The security machines are being
introduced across the country and create 'naked' images of
passengers</description>
</contentMeta>
contentSet>
remoteContent href="/images/ImageHandler.ashx?id=122771" size="49119"
height="396" width="468" contenttype="image:JPG" />
</contentSet>
</newsItem>
</packageItem>
</itemSet>
    
```

`</newsMessage>`

c. Appendix 3: What is NITF?

NITF or “News Industry Text Format” is a IPTC standard for representing and transmitting textual components of articles and is encapsulated within the NewsML-G2 packageItem as a newsItem.

d. Appendix 4: Website Title Codes

Website Title	Title Code	Publisher
blogs.FT.com	WEBFTB	The Financial Times Ltd.
Dailymail.co.uk	WEBDM	Associated Newspapers Ltd
Dailystar.co.uk	WEBDS	Northern and Shell Media Publications
Eveningexpress.co.uk	WEBEE	Aberdeen Journals Ltd.
Express.co.uk	WEBDE	Northern and Shell Media Publications
FT.com	WEBFT	The Financial Times Ltd.
Guardian.co.uk	WEBG	Guardian News and Media Ltd
Independent.co.uk	WEBI	Independent News & Media plc
Mirror.co.uk	WEBDMI	Trinity Mirror plc
Observer.guardian.co.uk	WEBO	Guardian News and Media Ltd
People.co.uk	WEBP	Trinity Mirror plc
Pressandjournal.co.uk	WEBPJ	Aberdeen Journals Ltd.
Telegraph.co.uk	WEBDT	Telegraph Media Group Limited
Thecottishsun.co.uk	WEBTSSUN	News International Ltd
Thesun.co.uk	WEBTSUN	News International Ltd
Thesundaytimes.co.uk	WEBSTO	News International Ltd
Thetimes.co.uk	WEBTTO	News International Ltd

e. Appendix 5: Search Noise Words

	Noise Word
A	a about after all also an and another any are

	as at
B	b be because been before being between both but by
C	c came can come could
D	d did do does
E	e each else
F	f for from
G	g get got
H	h has had he have her here him himself his how
I	i

	if in into is it its
J	j just
K	K
L	l like
M	m make many me might more most much must my
N	n never no now
O	o of on only or other our out over
P	p
Q	q
R	r re
S	s said same see

	<p>should since so some still such</p>
T	<p>t take than that the their them then there these they this those through to too</p>
U	<p>u under up use</p>
V	<p>v very</p>
W	<p>w want was way we well were what when where which while who will with would</p>

X	x
Y	y you your
Z	z
Numbers	1 2 3 4 5 6 7 8 9 0
Characters	\$