JRC.IPSC.G02 – EMM

Content Provider system interfaces

EP Media Monitor (EPMM) platform powered by EMM technology

Summary: This document provides the specification of the interfaces between the Press Review subsystem and the Content Providers.

Rel.: 1.0

Marco Verile, Aldo Podavini, Erik van der Goot

Table of Contents

Change log

Release	Date	Remarks
1.0	06/08/2015	Protocol release 1.0.

1 Overview

1.1 Purpose of this document

Technical notes summarizing key pieces of information driving design and implementation: detailed requirements, architectural and detail design, constraints, formats, etc.

1.2 Glossary

Term	Description
Summary	Text describing a piece of news.
Description	Generic term: it can contain a news summary.
News item	The metadata used to describe a piece of news.
Enclosures	any type of attachments
Cutting	Enclosure containing a PDF image from a newspaper or on-line page

Other resources:

- RSS 2.0 specification: http://cyber.law.harvard.edu/rss/rss.html
- Dublin Core: http://dublincore.org

1.3 Scope

The Press Review subsystem of the European Parliament Media Monitor platform addresses two main tasks:

1. the upload of news items and enclosures;
2. the editorial control of the reports (Daily Press Reviews, Audio Visual Reviews, etc.)

The tools provided to perform the above tasks are web-based applications:

1. NewsDesk (workspace): content management, report editing/publishing, etc.
2. Document Upload: upload of news items metadata and enclosures.

Figure 1. Document Upload GUI used to upload news items (metadata and enclosures).

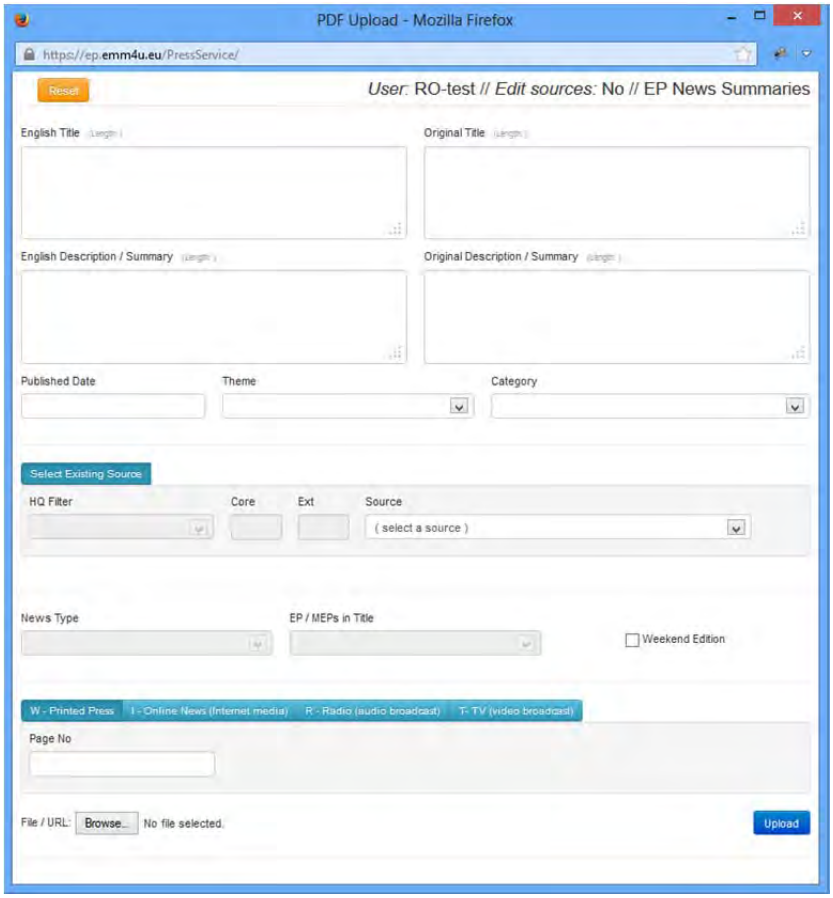
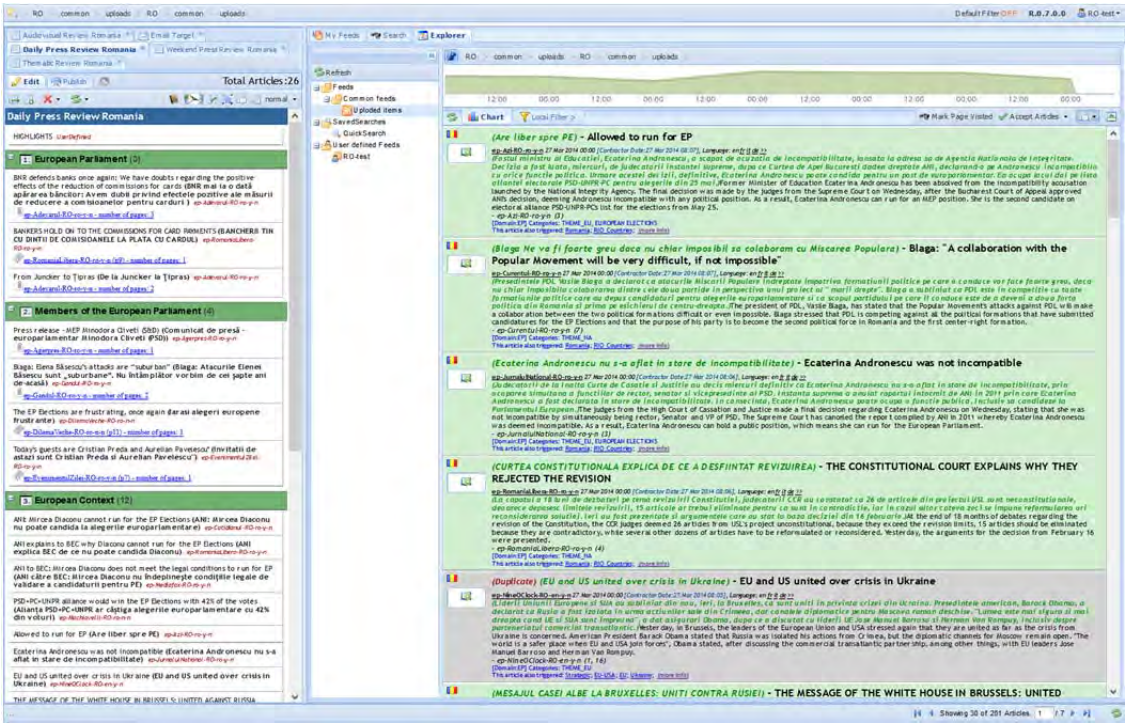


Figure 2. NewsDesk workspace (GUI) used to edit the final report (Daily Press Review).



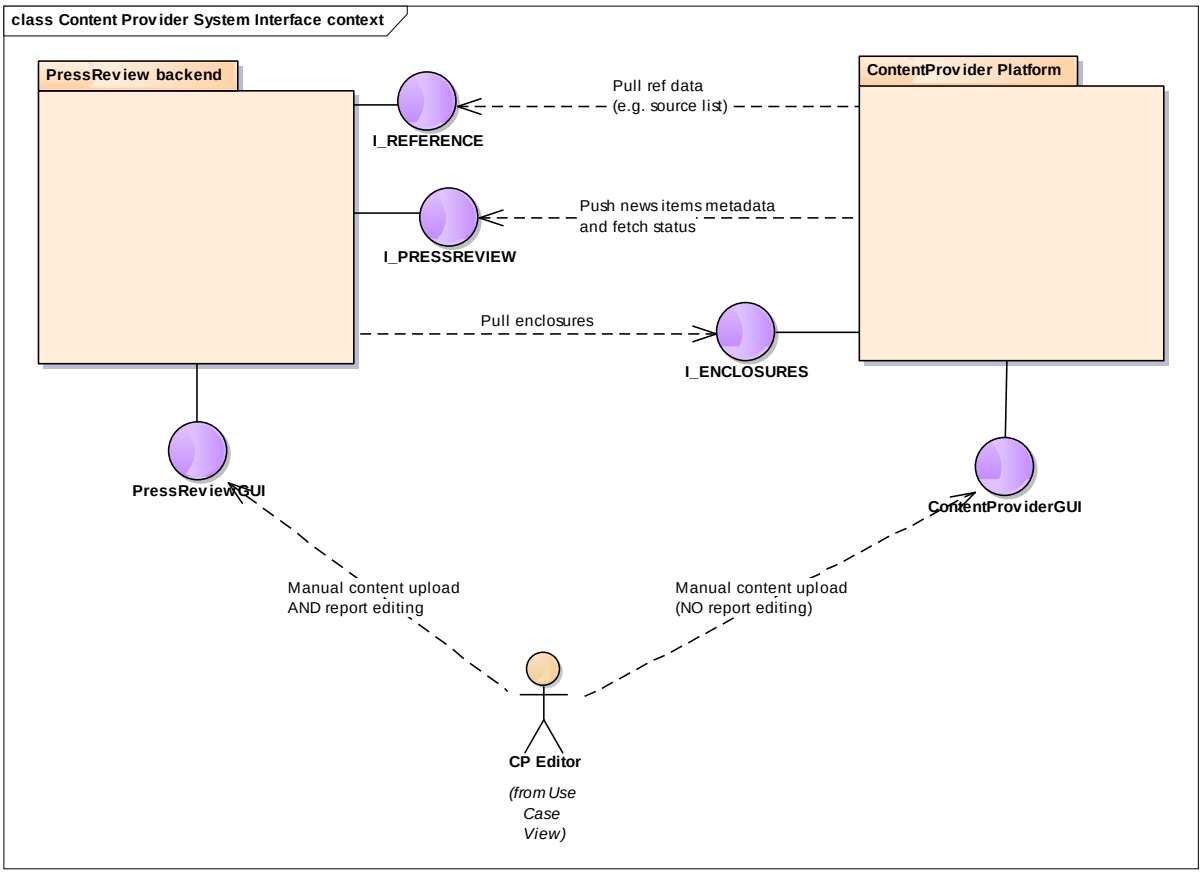
An alternative way to upload news items metadata and enclosures is based on the Press Review system interfaces (see Figure 3).

Both EPMM-Press Review system as well as the Content Provider platform shall provide users with a Graphical User Interface to upload metadata and enclosures. The two platforms shall exchange messages to make sure that news items metadata and related enclosures are delivered to the EPMM Press Review system.

Note that Content Provider editors shall anyway use the NewsDesk tool to complete the editing, and publication of the reports.

The GUIs are out of the scope of this document that focuses on the system interfaces only.

Figure 3. Context Diagram.



2 Reference metadata

2.1 Sources

Sources are agreed between national content providers and EP headquarters. The EP staff maintains the list of sources that can be downloaded in XML format from the Press Review platform:

- https://ep2.emm4u.eu/ContentProvider/reference/channeldirectory.xml

The source metadata relevant to the upload system interface includes the following fields:

Field	Description	Notes
channel/@id	Source unique identifier.	To be referenced within the news items metadata.
channel/ranking	Possible values are: • "**core**" : media analysis metadata is requested • "**non-core**": no media analysis metadata requested.	See "RULE 2: Media Analysis fields".
channel/dc:description	Display name.	To be displayed to the end-user (editor) using the Content Provider GUI.
channel/emm:storage	Possible values are: • "**external**" : enclosures will not be uploaded (as they need to be directly accessed via the source distribution platform). • "**internal**": enclosures will be uploaded and served by the EPMM platform to end-users.	In case of external storage, the EPMM platform will not download[1] the enclosures and will provide the URLs directly to the end-user.
channel/iso:country	Country ISO code of the Source.	
channel/iso:language	Language ISO code of the	It is used to determine the

[1] In this version of the specification the EPMM platform will not check whether the supplied URL is actually referring to an existing external resource.

	source	news articles original language.
channel/ocs:encoding	Expected character encoding.	

Sample sources metadata:

```
<directory xmlns="http://emm.jrc.it" xmlns:dc="http://purl.org/metadata/dublin_core#" xmlns:emm="http://emm.jrc.it"
xmlns:iso="http://www.iso.org/3166" xmlns:ocs="http://purl.org/ocs/directory/0.5/#"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
        <channel id="ep-KerrymanTraleeEdition-IE-en-n-y">
                <dc:description>Kerryman Tralee Edition</dc:description>
                <ocs:encoding>UTF-8</ocs:encoding>
                <iso:country>IE</iso:country>
                <ranking>non-core</ranking>
                <iso:language>en</iso:language>
                <emm:storage>internal</emm:storage >
        </channel>

        <channel id="ep-AXIA-GR-el-n-n">
                <dc:description>AXIA</dc:description>
                <ocs:encoding>UTF-8</ocs:encoding>
                <iso:country>GR</iso:country>
                <ranking>non-core</ranking>
                <iso:language>el</iso:language>
                <emm:storage>external</emm:storage >
        </channel>

 [ .... ]
</directory>
```

2.2 Categories

EP headquarters staff maintains a reference list of categories that can be downloaded in XML format from the Press Review platform:

- https://ep2.emm4u.eu/ContentProvider/reference/ep-categories.xml

Four categories are also called themes: THEME_EP, THEME_MEP, THEME_EU, THEME_NA and need always to be assigned to each news item.

Besides the theme, also one of the other categories is expected for every news item.

Sample categories:

```
<?xml version="1.0" encoding="UTF-8"?>
<categoryInfo xmlns:fn="http://www.w3.org/2005/xpath-functions" xmlns:xs="http://www.w3.org/2001/XMLSchema" startDate="Mon Mar 03
09:40:57 CET 2014" update="Thu Mar 13 15:41:12 CET 2014">
        <categories>
                <category domain="EP" id="THEME_EP">
                        <description>European Parliament</description>
                        <contact>florentina.ciltu@europarl.europa.eu</contact>
                        <class>theme</class>
                        <maxArticles>100</maxArticles>
                </category>
                <category domain="EP" id="THEME_MEP">
                        <description>Members of the European Parliament</description>
                        <contact>florentina.ciltu@europarl.europa.eu</contact>
                        <class>theme</class>
                        <maxArticles>100</maxArticles>
```

```
                    </category>
                    <category domain="EP" id="THEME_EU">
                            <description>European context</description>
                            <contact>florentina.ciltu@europarl.europa.eu</contact>
                            <class>theme</class>
                            <maxArticles>100</maxArticles>
                    </category>
                    <category domain="EP" id="THEME_NA">
                            <description>National context</description>
                            <contact>florentina.ciltu@europarl.europa.eu</contact>
                            <class>theme</class>
                            <maxArticles>100</maxArticles>
                    </category>
                    <category domain="EP" id="THEME_THEMATIC">
                            <description>Thematic Review</description>
                            <contact>florentina.ciltu@europarl.europa.eu</contact>
                            <class>theme</class>
                            <maxArticles>100</maxArticles>
                    </category>
                    <category domain="EP" id="AGRICULTURE_AND_RURAL_DEVELOPMENT">
                            <description>AGRICULTURE AND RURAL DEVELOPMENT</description>
                            <contact>florentina.ciltu@europarl.europa.eu</contact>
                            <class>theme</class>
                            <maxArticles>100</maxArticles>
                    </category>
                    <category domain="EP" id="BUDGETS">
                            <description>BUDGETS</description>
                            <contact>florentina.ciltu@europarl.europa.eu</contact>
                            <class>theme</class>
                            <maxArticles>100</maxArticles>
                    </category>
[ .... ]


            </categories>
</categoryInfo>
```

3 News Items Metadata

Selected news articles are described by metadata that shall be pushed to the EPMM Press Review system in RSS format.

Sample RSS with news items metadata:

```
<?xml version="1.0" encoding="UTF-8"?>
<rss version="2.0" xmlns:emm="http://emm.jrc.it" xmlns:iso="http://www.iso.org/3166">

<channel>
    <title>DPR Items for workgroup: RO</title>
    <pubDate>2014-01-20T03:16+0100</pubDate>

<item emm:newsType="article" emm:epInTitle="n" emm:edition="weekend" >
    <guid>af373f0bab987be983fd2a948efb7859</guid>

    <title>Public tenders attributed according to new criteria</title>
    <description>The European Parliament imposes a single legislation of public tenders and concession contracts in the European Union, according to a press release. The new rules ensure a bigger quality and a better price when public authorities buy or rent activities, goods or services. The new European legislation will facilitate access to tenders of small and medium companies and includes harsher provisions regarding subcontracting.</description>
    <iso:language>en</iso:language>

    <emm:title lang="ro">Licitatiile publice atribuite dupa noi criterii</emm:title>
    <emm:description lang="ro">Parlamentul European impune o legislatie unica a licitatiilor publice si a contractelor de concesionare in Uniunea Europena, potrivit unui comunicat de presa. Noile reguli asigura o calitate mai mare si un pret mai bun atunci cand autoritatile publice vor cumpara sau inchiria activitati, bunuri sau servicii. Noua legislatie comunitara va facilita accesul la licitatii al firmelor mici si mijlocii si include prevederi mai stricte privind subcontractarea.</emm:description>

    <pubDate>2014-01-17T00:00+0100</pubDate>
    <source >JurnalulNational-RO-ro-y-n</source>

    <enclosure emm:mediaType="W" emm:page="7" type="application/pdf" url="•
    https://contentproviderplatform.com/writtentpress/article2345.pdf" />

    <category domain="EP">THEME_EP</category>
    <category domain="EP">LEGAL_AFFAIRS</category>

    <emm:text >Figura 1: Public tenders attributed according to new criteria Ziar Autor Pagina Suprafata JURNALUL NATIONAL Elena Stan 7 312.80 1</emm:text>
</item>
```

```
[…]
</channel>
```

3.1 Detailed description

3.1.1 channel element

Field	Description	Notes
channel/title	Human readable text about the whole RSS file content.	Mandatory. Set by the content provider.
channel/pubDate	Publish date/time declared by the content provider.	Mandatory. Values shall conform to the following format (see 8.4): "EEE, dd MMM yyyy HH:mm:ss z" Example: “Thu, 04 Sep 2014 16:46:00 CEST”

3.1.2 item element

Field	Description	Notes
item/@emm:edition	Accepted values: • “weekday” • “weekend”	Mandatory. Set by the content provider.
item/@emm:epInTitle	Accepted values: "n" or ”y”.	Mandatory in case business “RULE 2: Media Analysis fields” is satisfied.
item/@emm:newsType	Accepted values: • "article", • "interview", • "editorial", • "in_brief", • "front_page", • "dossier", • "others".	Mandatory in case business “RULE 2: Media Analysis fields” is satisfied.
item/@emm:update	If present, it shall be set to “true”.	Optional Only in case the current item shall replace a previous version of the same item (same guid). Note: items can be replaced in the

		internal system but not in the products already created. If an item is re-sent with the guid of an existing item, the data consistency is not guaranteed.
item/source	Source identifier of the publisher.	Mandatory. The list of allowed source identifiers can be retrieved via the reference interface (see 2.12.1).
item/title	English title of the news item.	Mandatory.
item/description	English description of the news item (also referred as “summary” by editors)	Mandatory.
item/iso:language	Always “en”.	Mandatory. Always English for main title and description.
item/emm:title	Title in the original language (if not in English).	Not to be submitted if language (item/emm:title/@lang)is “en”. Otherwise Mandatory.
item/emm:title/@lang	Original language ISO code	Mandatory if item/emm:title is present.
item/emm:description	Original language description (if not in English).	Not to be submitted if language (item/emm:description/@lang)is “en”. Otherwise Mandatory.
item/emm:description/@lang	Original language description language ISO code	Mandatory if item/emm:description is present.
item/pubDate	Date at which the news item source actually published the article.	Mandatory. Values shall conform to the following format (see 8.4): "yyyy-MM-dd'T'HH:mmZ" Example: "2014-01-17T08:06+0100"
item/iso:language	“en”	Mandatory.
item/guid	Unique identifier set by the contractor in the	Mandatory.

	following format: cpID • cpID is a string (up to 32 characters long) only including the following characters: o lowercase letters (a-z) o 0-9 digits Examples: 0062444f0ad08e5aa94dfb4a383c6ac5 348	
item/emm:text	Extracted full text (OCR) / transcript of the news item.	Optional.
item/emm:text/@lang	Language ISO code of the *item/emm:text* element.	Mandatory if item/emm:text is present.

3.1.3 category element

Field	Description	Notes
item/category	Category Identifier as described in chapter 0. Two category elements are expected: 1 theme-category and 1 policy-area category.	Mandatory. 1. Special themes categories: • THEME_EP • THEME_MEP • THEME_EU • THEME_NA • THEME_THEMATIC 2. One additional non-theme among the others from the reference list (see 2.2).
item/category/@domain	Always set to "EP".	Mandatory.

3.1.4 enclosure element

Field	Description	Notes
enclosure/@emm:mediaType	Accepted values: • "W" for written press • "I" Internet page • "T" Television broadcast • "R" Radio broadcast	Mandatory.
enclosure/@emm:page	Page(s) where news article has been published.	Only in case of printed press (Enclosure/@emmmediaType = 'W')

	Examples: “2”, “4-5”, “1,6”.	
enclosure/@length	File length in bytes.	Mandatory only if the enclosure/@url is present.
enclosure/@type	Accepted values include (lowercase): • "audio/mpeg" , • “application/pdf”, • “video/mp4”.	Mandatory. Attachment file MIME type.
enclosure/@url	URL where the enclosure can be downloaded	Optional: depends on the contract clauses. In case of external sources the URL is directly used by the end user and not by the EPMM platform.
enclosure/@emm:broadcastTime	24h time of the day formatted as: HH:mm	Mandatory in case of enclosure/@emm:mediaType = “T” Examples: “08:15”, “20:30”.
enclosure/@emm:broadcastDuration	Duration in minutes. Format: positive integer. Examples: “3”, “12”, “90” minutes.	Mandatory in case of enclosure/@emm:mediaType = “T”
enclosure/@emm:broadcastTitle	Free text TV show / Radio broadcast title.	Optional.
enclosure/@emm:author	Free text containing author(s) names.	Optional.

4 Enclosures

Enclosures are files referred by news items. They can be PDFs, MP3, MP4 or other agreed formats. The content provider platform shall make those files available for download (https protocol with IP based authentication). Example:

- https://contentproviderplatform.com/writtenpress/article2345.pdf

Note: In case of source with external storage the enclosure will be directly downloaded by the end-user (providing a password if needed).

5 Business Rules

5.1 RULE 1: Mandatory fields

- English title
- Original title
- English description

- Published date
- Theme
- Source

5.2 RULE 2: Media Analysis fields

If **(theme-category is either THEME_EP or THEME_MEP) AND (source is core)** the following fields are mandatory:

- 'News type' (item/@ emm:newsType)
- 'EP/MEPs in Title' (item/@emm:epInTitle)

5.3 RULE3: enclosure with URLs (clippings)

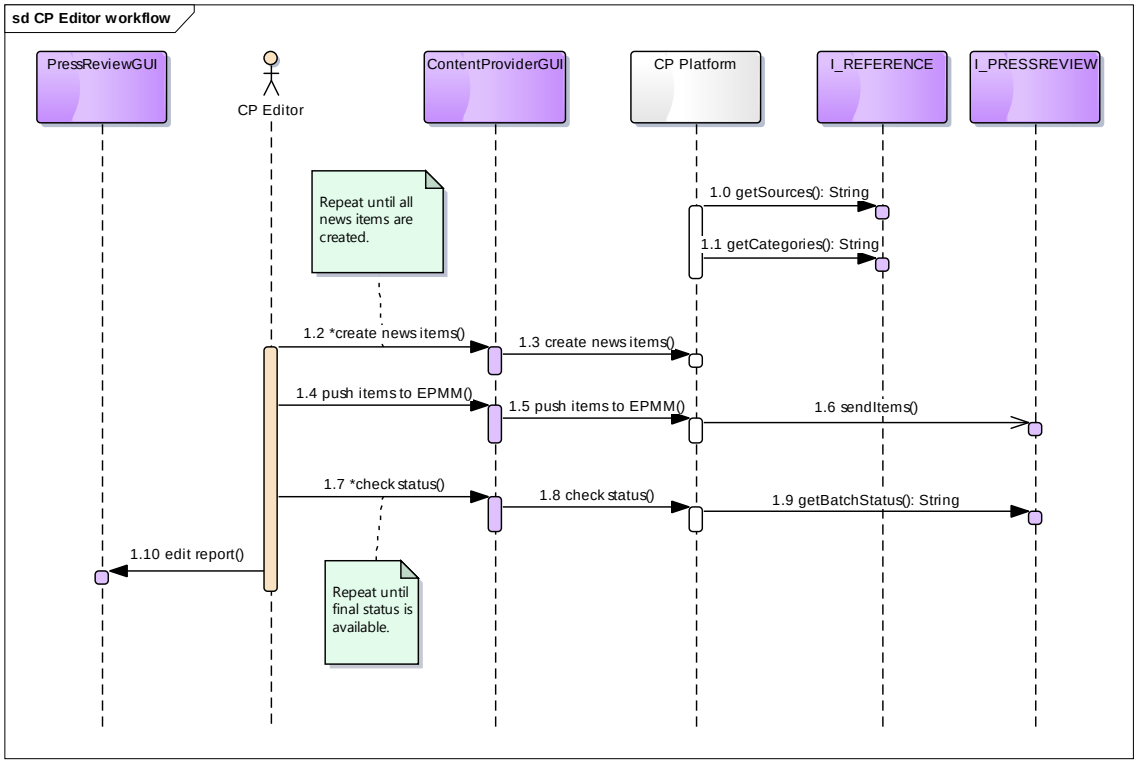
Enclosures elements shall have the corresponding url attribute only when **theme-category is either THEME_EP or THEME_MEP.**

6 Workflow

From the Content Provider Editors (CP Editor) perspective, the daily work is performed using two tools (See Figure 4) :

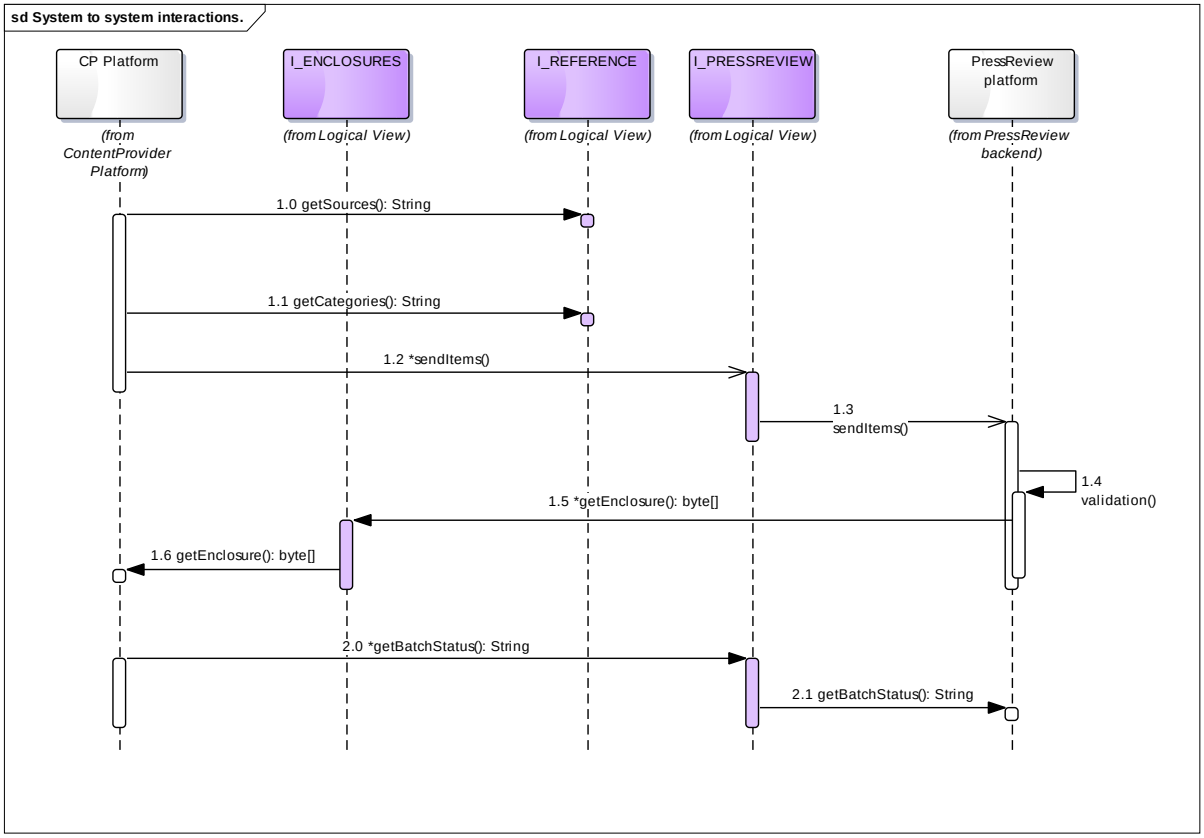
- By using the Content Provider GUI they will create the news items, push them to the EPMM platform and check whether the transactions successfully completed.
- With the EPMM PressReview GUI (NewsDesk workspace tool) they will finalize the editing of the report (all existing business rules apply).

Figure 4. CP Editor workflow.



The communication between the CP Platform and the PressReview platform is described in the following diagram:

Figure 5. CP Platform and PressReview system operations.



Two operations (getSources() and getCategories()) provide reference metadata. CP platform can call those operations anytime. It is responsibility of the CP platform to detect changes in the reference data[2].

Once the list of news items is ready, including all metadata and related enclosures, the CP platform can submit the metadata (sendItems() operation). PressReview will return a batch identifier to be used later on to retrieve the batch status.

The PressReview platform validates the metadata of each news item collecting independent success/error messages. If the link to download the enclosure is present it will also start downloading the related resource (getEnclosure() operation on the I_ENCLOSURE interface of the CP platform).

The CP platform can query the PressReview system to get the status of the submitted items (successfully processed or rejected).

[2] HTTP Last-Modified header can be used to know whether the remote file has been updated.

7 Batch Status

The ContentProvider platform can query the status of each submitted batch by using the getBatchStatus() operation. The status of each batch is described as follows:

7.1 batchInfo element

Field	Description	Notes
batchInfo/@id	Identifier returned by the sendItems() operation.	Empty in case of invalid id.
batchInfo/@submittedOn	Date/time at which the platform received the batch.	Empty in case of invalid id.
batchInfo/@status	Possible values: • batch_pending, • batch_completed, • batch_aborted, • invalid_batchId	• **batch_pending**: not yet completed (try later). • **batch_completed**: return details of the items. • **batch_aborted**: fatal error batch needs to be re-submitted. • **invalid_batchId**: batch id does not exist (never submitted). Resubmit correct id.
batchInfo/@statusMsg	Free text description.	Optional
batchInfo/@totItems	Total number of news items within the batch.	Only in case of status=batch_completed.
batchInfo/@rejectedItems	Total number of items rejected because of errors.	Only in case of status=batch_completed.

7.2 itemInfo element

These elements are present only in case of errors.

Field	Description	Notes
itemInfo/@guid	GUID.	
itemInfo /error	Error message.	If present, the item will be rejected.
itemInfo /error/@code	Error code[3].	

[3] List of error codes will be included in a later version of this specification.

7.3 Sample status xml data

```
<?xml version="1.0" encoding="UTF-8"?>
<batchInfo id="DF20342" submittedOn="2014-01-29T13:49+0100" status="batch_completed" totItems="34" rejectedItems="2">

    <itemInfo guid="ccp20203"  >
        <error code="105">Cannot download enclosure https://contentprov.com/mydata/345456967.pdf</error>
    </itemInfo>

    <itemInfo guid=" ccp202a5" >
        <error code="203">Invalid source identifier</error>
    </itemInfo >

</batchInfo>
```

8 API

Protocol: HTTPS

Authentication: to be agreed, preferably IP based authentication.

8.1 Interface I_REFERENCE

8.1.1 getSources

- Protocol: HTTP GET
- URL: https://ep.emm4u.eu/ContentProvider/reference/channeldirectory.xml
- Return: XML file as described in chapter 2.
- Remarks: In order to know whether the file has been updated, use the HTTP “Last-Modified” header.

8.1.2 getCategories

- Protocol: HTTP GET
- URL: https://ep.emm4u.eu/ContentProvider/reference/ep-categories.xml
- Return: XML file as described in chapter 2.
- Remarks: In order to know whether the file has been updated, use the “Last-Modified” header.

8.2 Interface I_PRESSREVIEW

8.2.1 sendItems

- Protocol: HTTP POST (content-type: multipart/form-data)
- URL: https://ep.emm4u.eu/ContentProvider/pressreview
- Parameters:

Name	Value	Notes
version	"1.0"	• Content-Type: text/plain • charset=UTF-8
action	"sendItems"	• Content-Type: text/plain • charset=UTF-8
rss	XML data	• Content-Type: application/octet-stream • Content-Transfer-Encoding: binary

- Return:
  - In case of success, the HTTP code 200 and a string representing the batch identifier.
  - In case of error, error-specific HTTP code (400, 500, etc..,) and a plain text error message.
- Remarks: none.

8.2.2 getBatchStatus

- Protocol: HTTP GET
- URL: https://ep.emm4u.eu/ContentProvider/pressreview
- Parameters:
  - version: "1.0"
  - action: getBatchStatus
  - batchId: batch identifier as returned by the sendItems operation.
- Return:
  - In case of success, the HTTP code 200 and an XML file with a status entry for each submitted news item.
  - In case of error, error-specific HTTP code (400, 500, etc...) and a plain text error message.

8.3 Interface I_ENCLOSURES

8.3.1 GetEnclosure

- Protocol: HTTP GET
- Call: https://contentproviderplatform.com/writtentpress/article2345.pdf

Return: the enclosure file.

8.4 Date formats

Date formats are expressed using the following pattern letters:

- EEE Day name in week (ex. “Tue”)
- dd Day in month (ex. “05”)
- MMM Month in year (ex. “Jul”)
- yyyy Year (ex. “2014”)
- HH Hour in day (0-23) (ex. “03”)
- mm Minute in hour (ex. “35”)
- ss Second in minute (ex. “55”)
- z General time zone (ex. “PST”, “CEST”)