JLE Legacy XML import

XML delivery requirement

dolomède

Confidential - Rév. 13 - 10/11/2010

Table of contents

1 FTP

XML imports are received on an FTP account on the jle.com hosting platform.

Test account :

- server: test.doloforge.com
- login / password: jle_jle4_import / (sent separately)
- directory: drop the dataset files directly inside the root directory

Production account :

- server: prod.doloforge.com or www.jle.com
- login / password: ditto “test”

XML files in the new format will be dropped into another FTP account (to be created).

Texte.

> Remarque

```
Code informatique
```

2 Dataset

The files sent on the server are tarball (*.tar.gz) archives, called "dataset"s. This format is just a way to send a single file, containing a full issue. Tarballs are extracted in the FTP directory before importing.

Once extracted, datasets are renamed, in order to be kept in-place, without being re-imported later: prefixed with "_", suffixed with ".N" (with N a unique number).

A dataset can contain a single article, or a full issue. AoP (Ahead of Print) and Web-Only are examples of single-article datasets.

2.1 Naming

Issue dataset naming:

```
lib-bdc-95-11-reguliere.tar.gz
lib-[journal_code]-[volume_number]-[issue_number]-[issue_type].tar.gz
```

Single-article (Ahead of Print, Web-Only) dataset naming:

```
aop-ejd-01001.tar.gz
aop-[journal_code]-[unique_article_ID].tar.gz
web-nro-00138.tar.gz
web-[journal_code]-[unique_article_ID].tar.gz
```

Once extracted the dataset MUST expose a directory tree starting with the same naming, ie.:

```
./lib-bdc-95-11-reguliere.tar.gz
```

extracts as:

```
./lib-bdc-95-11-reguliere/1013-1013-Breve-1/
./lib-bdc-95-11-reguliere/1013-1016-Breve-1/
./lib-bdc-95-11-reguliere/1016-1017-Breve-1/
./lib-bdc-95-11-reguliere/1021-1028-Article-1/
./lib-bdc-95-11-reguliere/1029-1038-Article-1/
./lib-bdc-95-11-reguliere/1039-1045-Article-1/
./lib-bdc-95-11-reguliere/1047-1051-Article-1/
./lib-bdc-95-11-reguliere/1053-1062-Article-1/
./lib-bdc-95-11-reguliere/1063-1066-Article-1/
```

where each part means:

```
./lib-[journal_code]-[volume_number]-[issue_number]-[issue_type]/
[page_start]-[page_end]-[type_of_document]-[order_of_article_in_page] /
```

A web-only article dataset:

```
web-nro-00191.tar.gz
```

extracts as (with same meaning for pages and type of document):

```
./web-nro-00191/001-002-Article-191/
```

An ahead of print dataset:

```
aop-hma-00523.tar.gz
```

extracts as (with no notion of page and type of document):

```
./aop-hma-00523/
```

2.2 Codes and values

[journal_code]

("SELECT code_revue, lib_revue FROM tbl_revue WHERE NOT is_super AND is_papier ORDER BY code_revue;"):

```
abc         Annales de Biologie Clinique
age         Annales de Gérontologie
agr         Cahiers Agricultures
bdc         Bulletin du Cancer
bic         Bulletin infirmier du Cancer
ecn         European Cytokine Network
ejd         European Journal of Dermatology
epd         Epileptic Disorders
epi         Epilepsies
ers         Environnement, Risques & Santé
hma         Hématologie
hpg         Hépato-Gastro
ipe         l©Information Psychiatrique
jpc         Journal de Pharmacie Clinique
lli         La Lettre de l©Internat
mca         MT Cardio
med         Médecine
met         Médecine thérapeutique
mrh         Magnesium Research
mtc         Médecine thérapeutique Cardiologie
mte         Médecine Thérapeutique Endocrinologie & Reproduction
mtg         MT / médecine de la reproduction, gynécologie et endocrinologie
mtm         Médecine Thérapeutique / médecine de la reproduction
mtp         Médecine thérapeutique / Pédiatrie
nro         Neurologie.com
nrp         Revue de neuropsychologie
ocl         Oléagineux, Corps Gras, Lipides
pnv         Psychologie & NeuroPsychiatrie du vieillissement
san         Cahiers d©études et de recherches francophones / Santé
sec         Science et changements planétaires / Sécheresse
sss         Sciences Sociales et Santé
stv         Sang Thrombose Vaisseaux
```

```
vir         Virologie
```

Some journal were renamed, leaving the old name there for technical reasons. Only the newer name must be used for new imports.

[issue_type]

(“SELECT code_type_parution, lib_type_parution FROM tbl_type_parution ORDER BY ordre, code_type_parution;”):

```
reguliere   Régulière
hors_serie  Hors série
num_double  Numéro double
num_special Numéro spécial
num_special2      Numéro spécial 2
supplement  Supplément
supplement_fmc    Supplément FMC
fmc         FMC
aop         Ahead of Print
web         Web Only
```

“reguliere” are normal issues; “aop” is automatically selected when importing an “aop-*” dataset; ditto for “web”. Other issue type are not technically defined, even though some may have special behavior once on the website. Details are known on the JLE side.

[page_start]-[page_end]

Full numeric values for the starting/ending page of the article.

[type_of_document]

The following documents are currently imported. Others are rather created using the website back-office. This may be improved in the future.

```
Article     regular printed (or aop) article
Breve       simple article with no DOI, no table of contents entry, etc.
```

Code is lowercased during import, so “Article” is the same as “article” (which is the official code).

[order_of_article_in_page]

Articles must have a unique [starting page] + [ending page] + [order in page]. This means that 2 articles in the same page must have a different value for [order_of_article_in_page] (ex. “1” and “2” resp.)

2.3 Issue

An issue is created in the website database, when an article is imported, and it's corresponding issue does not already exist. The details of the issue are taken from the dataset directory names (see above), instead of any XML values.

The “issue directory” should contain the first page of the issue, as a GIF image: “lib-bdc-95-11-reguliere/sommaire_parution.gif”

This MUST be in GIF format, 113 pixels width, approximately 146 to 166 pixels height (depending on the actual journal size). This image is displayed as-is on some pages, and also constrained to 90 pixels width on other pages (browser-side resizing).

The name of the file is obviously mandatory.

2.4 Article

An article (whatever type) is represented as a subdirectory within a dataset (ex. “./lib-bdc-95-11-reguliere/1013-1013-Breve-1/”, “./web-nro-00191/001-002-Article-191/ ”, “./aop-hma-00523/”), which contains the following files:

```
index.xml   the main XML file describing the article
index.htm   the XHTML/HTML file containing the main part of the article
images.htm  the XHTML/HTML file with the images/charts part of the article
index.pdf   the PDF version of the article (formatted for print)
jlebdc01045-gr1.jpg (various images linked from ªimages.htmº)
jlebdc01045-gr2.jpg
jlebdc01045-gr3.jpg
jlebdc01045-gr4.jpg
jlebdc01045-gr5.jpg
jlebdc01045-gr6.jpg
```

Other files:

```
*.flv       video contents (along with a special HTML tag)
lien_*      add-on contents (along with ad-hoc HTML links)
patient.htm version of the full-text for patients only (journal ªnroº)
```

XML

See samples in the existing datasets (*.tar.gz). Important notice: XML files are tidied-up during the import phase (nota: “JLE::Document::Import::XML::Jouve->decoupe_fichier”), which means the index.xml file residing on the FTP tree are NOT the ones originally sent. For complete reference, of the INPUT files, extract from the *.tar.gz.

The main fix is that input files are encoded in Latin1 with HTML entities, even though the XML file is declared in UTF-8. This is a legacy bug of the input files, and should not be changed.

Original XML version below (notice the “&#xc9;” used instead of “É”):

```
<?xml version="1.0" encoding="utf-8"?><!DOCTYPE Document SYSTEM "jle.dtd">
<Document Code="code" Nom="nom">
 <References>
  <DC.Title>&#xc9;ditorial</DC.Title>
  <DC.Language>fr</DC.Language>
  <Pages>
   <Edition.Page.Debut>1027</Edition.Page.Debut>
   <Edition.Page.Fin>1027</Edition.Page.Fin>
   <Edition.Page.Ordre>1</Edition.Page.Ordre>
  </Pages>
  <Edition.Author>D Orbach, C Massard, J-O Bay </Edition.Author>
<JLE.DOI>10.1684/bdc.2010.1184</JLE.DOI>
  <JLE.Affiliation></JLE.Affiliation>
  <JLE.DatePubli>2010-09-01</JLE.DatePubli>
  <JLE.DateParu>2010-09-01</JLE.DateParu>
  <JLE.Lib.Paru>septembre 2010</JLE.Lib.Paru>
  <JLE.Gratuit>0</JLE.Gratuit>
  <JLE.Lib.Somm>&#xc9;ditorial</JLE.Lib.Somm>
 </References>
 <Corpus></Corpus>
</Document>
```

Internally transformed version:

```
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE Document SYSTEM "/srv/doloforge/shared/jle/jle4/ftp/xml/jle.dtd">
<Document Code="code" Nom="nom">
  <References>
    <DC.Language>fr</DC.Language>
    <DC.Title>Éditorial</DC.Title>
    <Pages>
      <Edition.Page.Debut>1027</Edition.Page.Debut>
      <Edition.Page.Fin>1027</Edition.Page.Fin>
      <Edition.Page.Ordre>1</Edition.Page.Ordre>
    </Pages>
    <Edition.Author>D Orbach, C Massard, J-O Bay </Edition.Author>
    <JLE.DOI>10.1684/bdc.2010.1184</JLE.DOI>
    <JLE.Affiliation></JLE.Affiliation>
    <JLE.DatePubli>2010-09-01</JLE.DatePubli>
    <JLE.DateParu>2010-09-01</JLE.DateParu>
    <JLE.Lib.Paru>septembre 2010</JLE.Lib.Paru>
    <JLE.Gratuit>0</JLE.Gratuit>
    <JLE.Lib.Somm>Éditorial</JLE.Lib.Somm>
  </References>
  <Corpus></Corpus>
</Document>
```

Here are details about each element the importer cares about (not necessarily present in all examples, which in turn do not necessarily contain only things cared about). Element are denoted in XPath-like notation.

- Document: parameters are ignored, as the type of document and other details are extracted from the datset naming
- References/DC.Title: main title of the document, in vernacular language (exact rules are defined by JLE)
- References/JLE.DOI: DOI for the article; an new import for an existing DOI removes the

previously imported document, and replaces it with this new copy (Ahead of Print articles are tracked with their DOI)

- References/DC.Language: identifier of the language in which the article is written (“fr” or “en”)
- References/JLE.DatePubli: official date of publishing of the article, in ISO format (used for sorting)
- References/JLE.DateParu: official date of publishing of the issue, in ISO format (used for sorting, may be different than the previous date)
- References/JLE.Lib.Paru: full label for the issue, used for displaying (usually a simplified text version of the ISO date)
- References/JLE.Lib.Somm: label of the table of contents (“sommaire” in French) entry under which the article is available (the table of contents is reconstructed on the website, from the page order of the articles, grouped when they have the same “JLE.Lib.Somm”)
- References/Pages/Edition.Page.Debut: page number of the beginning of the article (integer) (those elements are copies of the values used in the dataset naming; exact value used is not completely clear)
- References/Pages/Edition.Page.Fin: page number of the end of the article (integer)
- References/Pages/Edition.Page.Ordre: order of the article in the starting page, if there are more than one (put “1” in case there is only one article starting on this page)
- References/JLE.Gratuit: 0 or 1, telling the website that this article is accessible free of charge (the info comes from JLE)
- References/JLE.Affiliation: single long text displaying the affiliation of the authors (HTML MAY be accepted)
- References/JLE.Reference: single long text displaying the references of the article (HTML MAY be accepted) (this tag is not used since a very long time, as references are now put inside the HTML for the article)
- References/JLE.References: as many elements of this kind, containing sub elements named “url”, “libelle”, “description”, “target”, “ordre” (this tag is not used since a very long time, as references are now put inside the HTML for the article)
- References/Edition.Author: single long text displaying the authors of the article (HTML MAY be accepted)
- CaracteristiquesFR/DC.Title: French version of the title, if available (may be the copy of the main “References/DC.Title” element), used on French pages on the website
- CaracteristiquesFR/Edition.Chapeau: not used anymore (french version)
- CaracteristiquesFR/Edition.Motscles: comma-separated list of French keywords
- CaracteristiquesFR/Edition.Resume: French summary of the article, which MAY contain HTML tags
- CaracteristiquesEN/DC.Title: English version of the title, if available (may be the copy of the main “References/DC.Title” element), used on English pages on the website
- CaracteristiquesEN/Edition.Chapeau: not used anymore (english version)

- CaracteristiquesEN/Edition.Motscles: comma-separated list of E,glish keywords
- CaracteristiquesEN/Edition.Resume: English summary of the article, which MAY contain HTML tags
- References/NEURO.Highlights: specifics for “Neurologie.com” journal
- References/NEURO.Classification: specifics for “Neurologie.com” journal
- References/NEURO.Lexique: specifics for “Neurologie.com” journal
- References/NEURO.HTMLPro: specifics for “Neurologie.com” journal
- References/NEURO.HTMLPatient: specifics for “Neurologie.com” journal

XM sample

Here is a more detailed XML sample:

```
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE Document SYSTEM "/srv/doloforge/shared/jle/jle4/ftp/xml/jle.dtd">
<Document Code="code" Nom="nom">
  <References>
    <DC.Language>fr</DC.Language>
    <DC.Title>Ostéonécroses et thromboses veineuses multiples : un cas de
patient porteur d'une nouvelle mutation de la protéine C (N102S) et
hétérozygote pour le facteur V Leiden</DC.Title>
    <Pages>
      <Edition.Page.Debut>437</Edition.Page.Debut>
      <Edition.Page.Fin>440</Edition.Page.Fin>
      <Edition.Page.Ordre>1</Edition.Page.Ordre>
    </Pages>
    <Edition.Author>M Benbih, E de Maistre, T Lecompte, D Mainard, M-C
Laprevote, M Alhenc-Gelas, J Devignes </Edition.Author>
    <JLE.DOI>10.1684/abc.2008.0234</JLE.DOI>
    <JLE.Affiliation>Laboratoire de biologie médicale, Centre hospitalier
de Chalons en Champagne, ... Paris</JLE.Affiliation>
    <JLE.DatePubli>2008-07-01</JLE.DatePubli>
    <JLE.DateParu>2008-07-01</JLE.DateParu>
    <JLE.Lib.Paru>Juillet-Août 2008</JLE.Lib.Paru>
    <JLE.Gratuit>0</JLE.Gratuit>
    <JLE.Lib.Somm>pratique quotidienne</JLE.Lib.Somm>
  </References>
  <CaracteristiquesFR>
    <Edition.Resume>L'association ... C circulante.</Edition.Resume>
    <Edition.Motscles>thrombose veineuse, ostéonécrose, thrombophilie,
déficit en protéine C, facteur V Leiden</Edition.Motscles>
  </CaracteristiquesFR>
  <CaracteristiquesEN>
    <DC.Title>Multiple osteonecroses and venous thrombosis: one case of
patient with a novel mutation of protein C (N102S) and heterozygous for FV
Leiden</DC.Title>
    <Edition.Resume>The association ... clinical studies.</Edition.Resume>
    <Edition.Motscles>venous thrombosis, osteonecrosis, protein C
deficiency, thrombophilia, factor V Leiden</Edition.Motscles>
  </CaracteristiquesEN>
  <Corpus></Corpus>
</Document>
```

HTML

The part within the <body> element is taken as the full-text of the article. Links to images must be simple (<img src="jlebdc01045-gr2.jpg">). Such links are parsed and file names are grabbed from them, which means that image file name may be arbitrary, as long as they reside in the same directory.

The index.htm file MAY contain links to the images.htm file, written as <a href="images.htm">. A link is added anyway at the top of the website page.

HTML text *per se* is simple, and displayed using a proper CSS. Some special tags may be needed to display Flash videos, etc.

The “images.htm” file is not mandatory. It is mainly designed to split large texts and large images. Details are not technical.

PDF

The PDF version of the article (a single PDF file) MUST be “index.pdf” (lowercase). File size is in the order of 100k (usually 50kB to 500kB).