Please clean my data ☺

Two scenarios

1. **Cleaning harvested metadata (Dublin Core, MARC, EAC)**
2. **Cleaning OCR text**

Vinita Tuteja
National Library of Australia

Cleaning harvested data

WHY

- Dirty data quite prominent in delivery systems

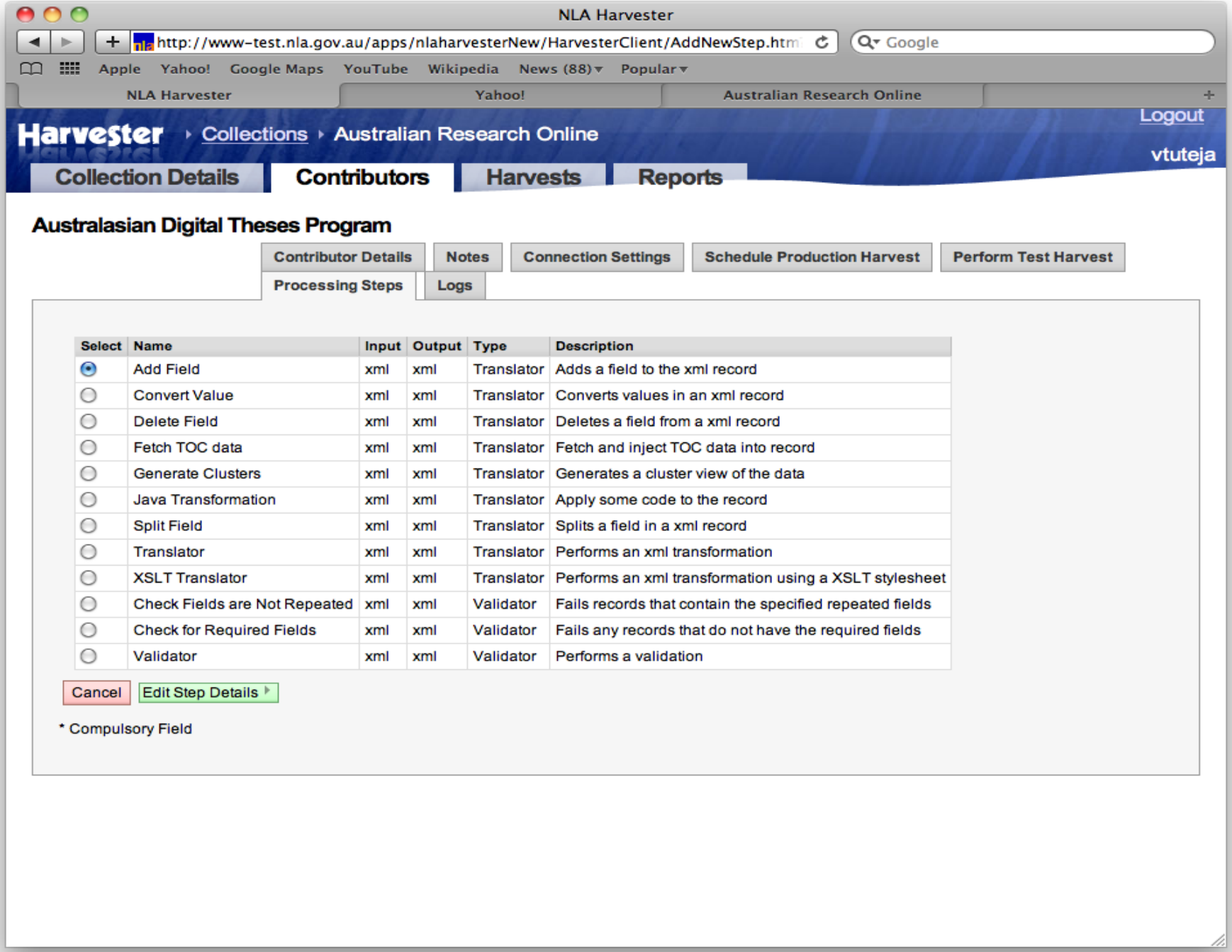
WHAT

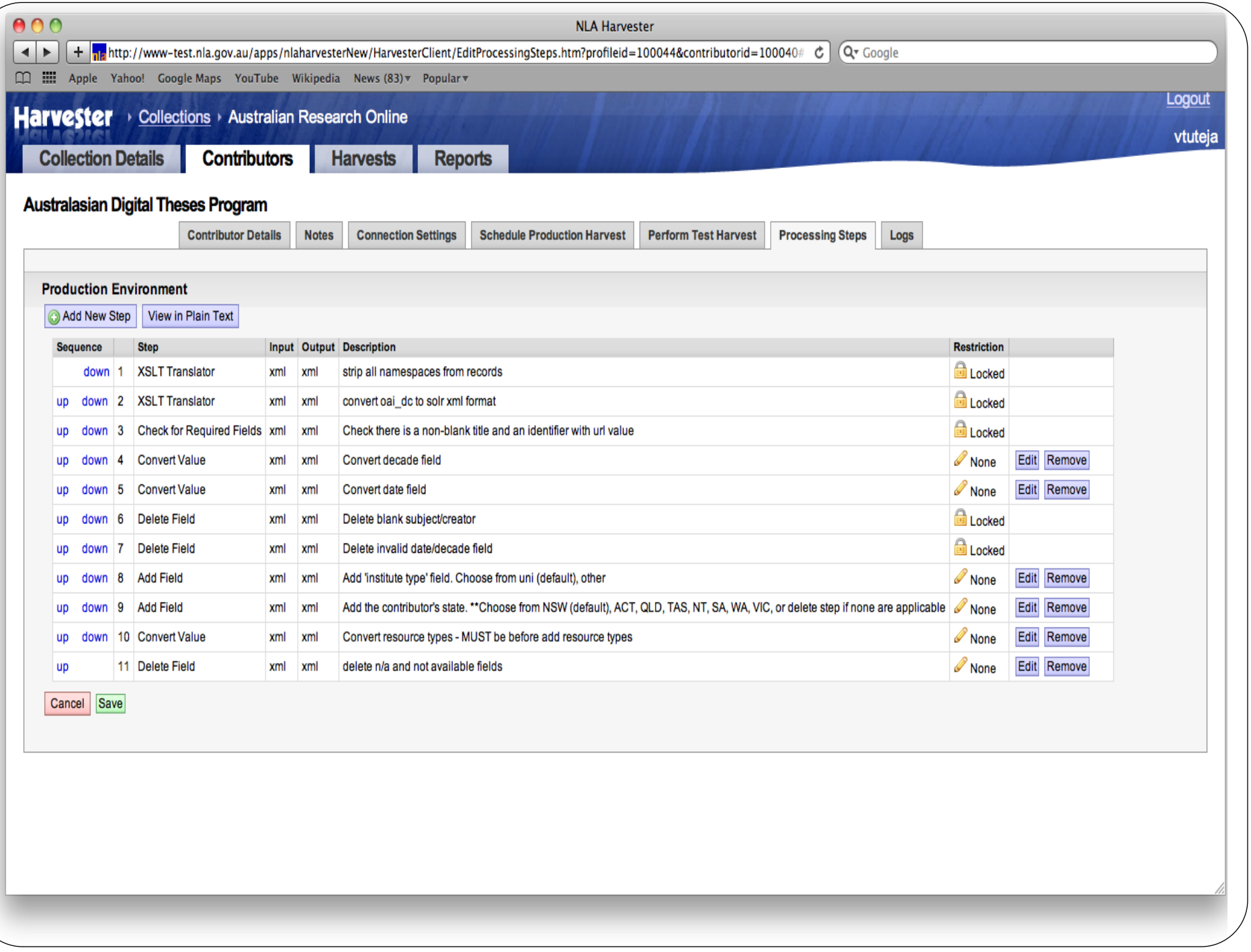
- Required tools to batch clean in a generic manner
- Should be simple enough for business owners to DIY

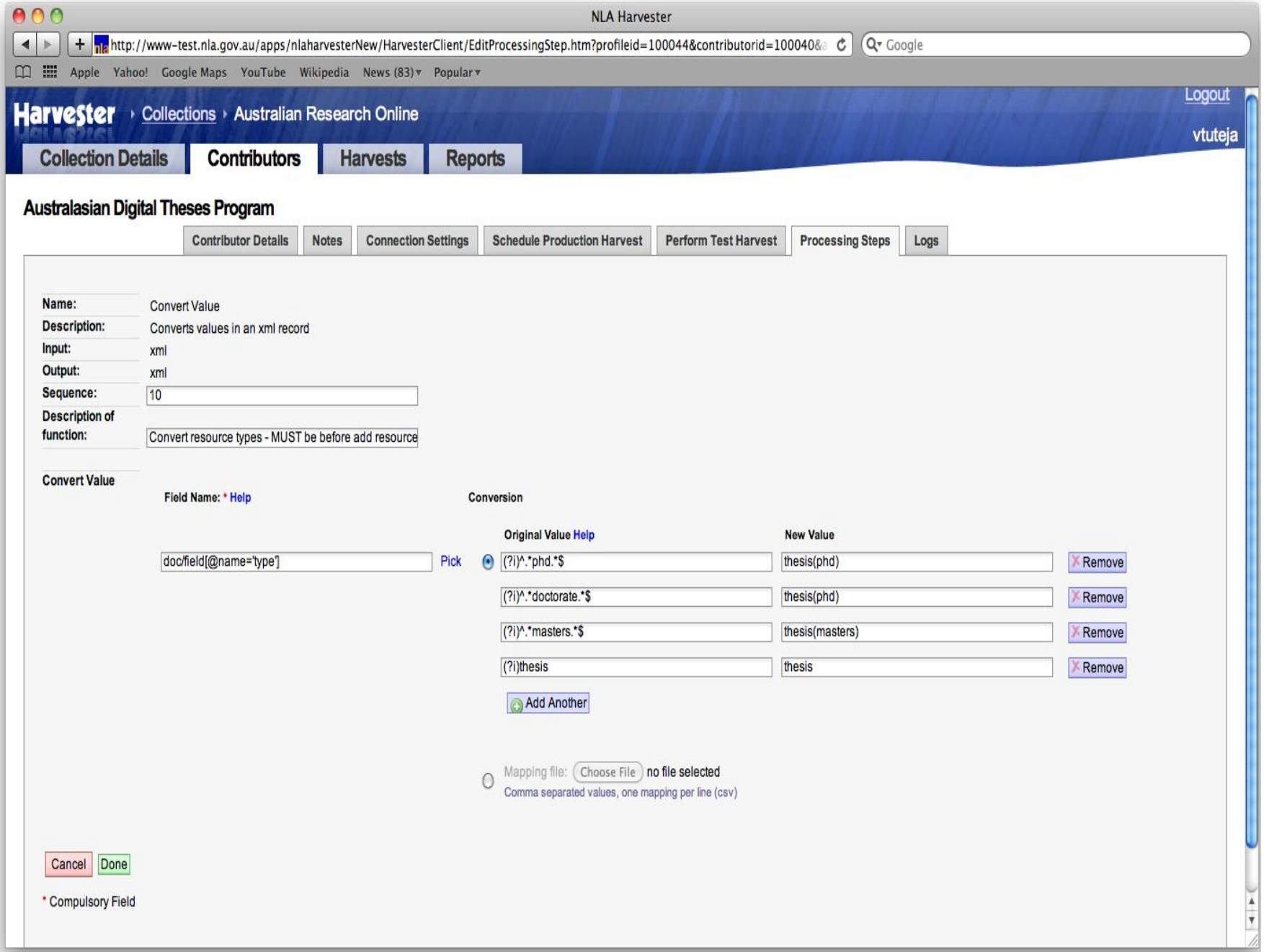
HOW

- Transformation and translation steps added to the harvesting process (all metadata records are in xml)
- Power of regular expressions made available as step templates e.g. Convert a value, delete a field, translate a step, add a new field …..
- Velocity Template variables used to store each regular expression match.

Using Velocity template variables gives us direct access to the xml elements (e.g. internally we store the xml document and matched element in $D and $E variables respectively). We then use the Java DOM4J API to do effectively whatever we wish.Velocity templates allow loops, if statements, variables, etc etc etc.


NLA Harvester
http://www-test.nla.gov.au/apps/nlaharvesterNew/HarvesterClient/AddNewStep.htm
Google
Apple Yahoo! Google Maps YouTube Wikipedia News (88) Popular
NLA Harvester
Yahoo!
Australian Research Online
Harvester
Collections
Australian Research Online
Logout
vtuteja
Collection Details
Contributors
Harvests
Reports
Australasian Digital Theses Program
Contributor Details
Notes
Connection Settings
Schedule Production Harvest
Perform Test Harvest
Processing Steps
Logs
Select
Name
Input
Output
Type
Description
Add Field xml xml Translator Adds a field to the xml record
Convert Value xml xml Translator Converts values in an xml record
Delete Field xml xml Translator Deletes a field from a xml record
Fetch TOC data xml xml Translator Fetch and inject TOC data into record
Generate Clusters xml xml Translator Generates a cluster view of the data
Java Transformation xml xml Translator Apply some code to the record
Split Field xml xml Translator Splits a field in a xml record
Translator xml xml Translator Performs an xml transformation
XSLT Translator xml xml Translator Performs an xml transformation using a XSLT stylesheet
Check Fields are Not Repeated xml xml Validator Fails records that contain the specified repeated fields
Check for Required Fields xml xml Validator Fails any records that do not have the required fields
Validator xml xml Validator Performs a validation
Cancel
Edit Step Details
* Compulsory Field


NLA Harvester
http://www-test.nla.gov.au/apps/nlaharvesterNew/HarvesterClient/EditProcessingSteps.htm?profileid=100044&contributorid=100040#
Google
Apple Yahoo! Google Maps YouTube Wikipedia News (83) Popular
Harvester
Collections
Australian Research Online
Logout
vtuteja
Collection Details
Contributors
Harvests
Reports
Australasian Digital Theses Program
Contributor Details
Notes
Connection Settings
Schedule Production Harvest
Perform Test Harvest
Processing Steps
Logs
Production Environment
Add New Step
View in Plain Text
Sequence
Step
Input
Output
Description
Restriction
down 1 XSLT Translator xml xml strip all namespaces from records Locked
up down 2 XSLT Translator xml xml convert oai_dc to solr xml format Locked
up down 3 Check for Required Fields xml xml Check there is a non-blank title and an identifier with url value Locked
up down 4 Convert Value xml xml Convert decade field None Edit Remove
up down 5 Convert Value xml xml Convert date field None Edit Remove
up down 6 Delete Field xml xml Delete blank subject/creator Locked
up down 7 Delete Field xml xml Delete invalid date/decade field Locked
up down 8 Add Field xml xml Add 'institute type' field. Choose from uni (default), other None Edit Remove
up down 9 Add Field xml xml Add the contributor's state. **Choose from NSW (default), ACT, QLD, TAS, NT, SA, WA, VIC, or delete step if none are applicable None Edit Remove
up down 10 Convert Value xml xml Convert resource types - MUST be before add resource types None Edit Remove
up 11 Delete Field xml xml delete n/a and not available fields None Edit Remove
Cancel
Save


NLA Harvester
http://www-test.nla.gov.au/apps/nlaharvesterNew/HarvesterClient/EditProcessingStep.htm?profileid=100044&contributorid=100040&
Google
Apple Yahoo! Google Maps YouTube Wikipedia News (83) Popular
Harvester
Collections
Australian Research Online
Logout
vtuteja
Collection Details
Contributors
Harvests
Reports
Australasian Digital Theses Program
Contributor Details
Notes
Connection Settings
Schedule Production Harvest
Perform Test Harvest
Processing Steps
Logs
Name: Convert Value
Description: Converts values in an xml record
Input: xml
Output: xml
Sequence: 10
Description of function: Convert resource types - MUST be before add resource
Convert Value
Field Name: * Help
Conversion
Original Value Help
New Value
doc/field[@name='type']
Pick
(?i)^.*phd.*$
thesis(phd)
Remove
(?i)^.*doctorate.*$
thesis(phd)
Remove
(?i)^.*masters.*$
thesis(masters)
Remove
(?i)thesis
thesis
Remove
Add Another
Mapping file:
Choose File
no file selected
Comma separated values, one mapping per line (csv)
Cancel
Done
* Compulsory Field

Summary

Status:	Stopped
Environment:	Test
Total records read:	404
Record updates read:	404
	View Records | View Records as XML | Download All Records
Records updates rejected:	10 (2.48%)
Record Errors:	

Stage	Error	Record Count
Check for Required Fields	Failed required field check on field doc/field[@name='title']	1
Check for Required Fields	Failed required field check on field doc/field[@name='identifier']	9

Log

UTC Time	Message	Attached Data
03.08.2009 03:52:08	Beginning Harvest [Local Time: 03.08.2009 13:52:08]	
03.08.2009 03:52:08	Fetching: http://metasuite.nun.unsw.edu.au:9004/CMD/OAI?verb=ListRecords&metadataPrefix=oai_dc&set=4E4F542852454C2853573B44432E4964656E7469666965723B4E42443A2929&from=2008-11-25T23:21:42Z	
03.08.2009 03:52:12	Harvested 101 records so far... [Local Time: 03.08.2009 13:52:12]	
03.08.2009 03:52:15	Record 8 rejected Processing Step: 3 . Check for Required Fields Reason: Failed required field check on field doc/field[@name='identifier'] OAI ID: 201968	View Data
03.08.2009 03:52:19	Fetching: http://metasuite.nun.unsw.edu.au:9004/CMD/OAI?verb=ListRecords&resumptionToken=202444%7C1227655302000%7C%7Coai_dc%7C4E4F542852454C2853573B44432E4964656E7469666965723B4E42443A2929	
03.08.2009 03:52:21	Harvested 202 records so far... [Local Time: 03.08.2009 13:52:21]	
03.08.2009 03:52:23	Record 35 rejected Processing Step: 3 . Check for Required Fields Reason: Failed required field check on field doc/field[@name='title'] OAI ID: 202483	View Data
03.08.2009 03:52:23	Record 45 rejected Processing Step: 3 . Check for Required Fields Reason: Failed required field check on field doc/field[@name='identifier'] OAI ID: 202493	View Data
03.08.2009 03:52:26	Fetching: http://metasuite.nun.unsw.edu.au:9004/CMD/OAI?verb=ListRecords&resumptionToken=202620%7C1227655302000%7C%7Coai_dc%7C4E4F542852454C2853573B44432E4964656E7469666965723B4E42443A2929	
03.08.2009 03:52:28	Harvested 303 records so far... [Local Time: 03.08.2009 13:52:28]	
03.08.2009 03:52:29	Record 11 rejected Processing Step: 3 . Check for Required Fields Reason: Failed required field check on field doc/field[@name='identifier'] OAI ID: 203437	View Data

Cleaning Newspapers OCR Text

- We empowered our users to correct OCR data for us and it took a life of its own
- We have a wonderful community built around this feature– diehard text correctors who compete with each other to feature in our top text correctors list
- They add value to the content by correcting the OCR, creating tags or leaving their comments.
- This has improved our indexes and hence the discovery of an article
- They create their own trails of discovery by using tags and make it available for others to follow
- WARNING!! This stuff is quite addictive.

Australia
Trove
one search ... a wealth of information

Login / Signup

All | Books, Journals, Magazines, Articles... | Pictures and Photos | Australian Newspapers (1803-1954) | Diaries, Letters, Archives... | Maps | Music, Sound and Video | Archived Websites (1996-now) | About People and Organisations

Newspapers: Home Browse Help

Historic Australian Newspapers, 1803 to 1954

Find an article

Search articles

Advanced Search

Find an Issue

by Title

1. The Sydney Morning Herald
2. The Argus
3. The Mercury (Hobart)
4. Northern Territory Times and Gazette
5. Brisbane Courier

Show all titles

by Date

1803

JAN	FEB	MAR	APR
MAY	JUN	JUL	AUG
SEP	OCT	NOV	DEC

S	M	T	W	T	F	S

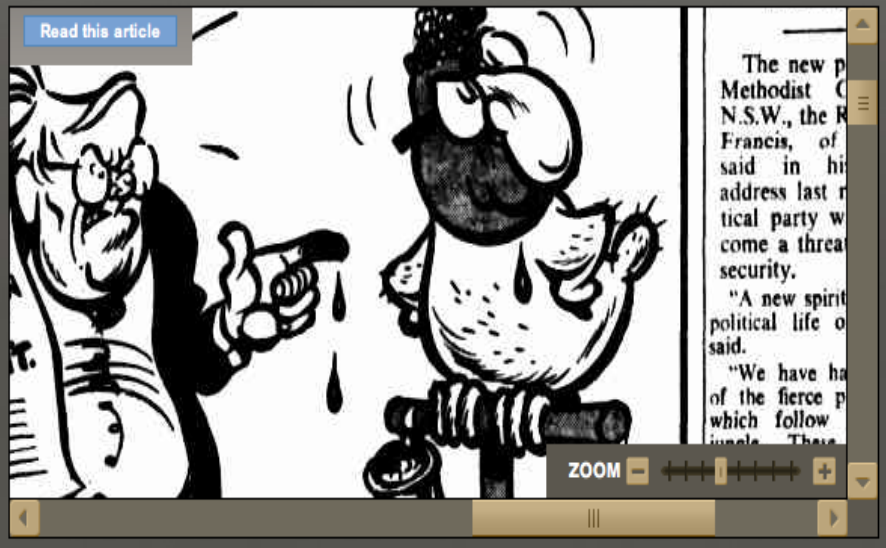
by State

On this day

THE SYDNEY MORNING HERALD (NSW : 1842-19..., WEDNESDAY 25 FEBRUARY 1953

Navigation tips for the example newspaper page below:

- **Scroll** with the scrollbars or your scrollwheel.
- **Pan** by clicking and dragging the image.
- **Zoom** with the zoom controls in the bottom right.



Top Text-Correctors

1. jhempenstall (364740)
2. John.F.Hall (337656)
3. fwalker13 (303448)
4. annmanley (273112)
5. maurielyn (245182)

Recent Comments

"Wamboota" in the text should be Womboot...
created 2010-02-25 11:51:43.0 by Corio

Total article editted.
created 2010-02-24 21:48:15.0 by Bazza59

Recent Tags

Hoare James James HOARE Elisha James Baynton Marian Heathcot Lucas

>> All tags

Newspaper Article

http://trove.nla.gov.au/ndp/del/article/11367711?searchTerm=world%20war%202

Google

Apple Yahoo! Google Maps YouTube Wikipedia News (85) Popular

Newspaper Article | Home – Trove

NATIONAL LIBRARY OF AUSTRALIA

Australia **Trove**

one search ... a wealth of information

Home | Tags | About | Site news | Contact us | Feedback

Login / Signup

All | Books, Journals, Magazines, Articles... | Pictures and Photos | Australian Newspapers (1803-1954) | Diaries, Letters, Archives... | Maps | Music, Sound and Video | Archived Websites (1996-now) | About People and Organisations

Newspapers: Home Browse Help

Search articles

Advanced Search

The Argus (Melbourne, Vic. : 1848-1954) (about) ◀ Saturday 28 October 1944 ▶ ◀ Page 4 S of 36 ▶

Print | Save as PDF | Save as Image

Cite: http://nla.gov.au/nla.news-article11367711

Tags (Keywords) Add New Tags

Comments Add New Comment

No comments yet.

ELECTRONICALLY TRANSLATED TEXT Fix this Text

Why may this text have mistakes?
How to correct this text?

No corrections yet

THE ROMANCE OF FLYING, No 56.
FLYING ACES OF WORLD WAR I
2.-Two VC Winners of 1915.
A QUARTER-CENTURY before "the Few" were to win immortal fame in 1940, war in the air was attracting a gallant company into the RFC. Like the first reconnaissance flights, the earliest fighting was more thrilling than useful. Airmen fought with anything they could lay their hands on - pistols, rifles, or even hand grenades., With improved methods of air-fighting, the RFC rapidly

View entire page

THE ROMANCE OF FLYING, No 56.

FLYING ACES OF WORLD WAR I

2.—Two VC Winners of 1915.

A QUARTER-CENTURY before "the Few" were to win immortal fame in 1940, war in the air was attracting a gallant company into the RFC. Like the first reconnaissance flights, the earliest

cate in 1911, was flying a BE b: plane, which carried a 100lb bomb. During the flight to Courtrai he met heavy machinegun and rifle fire. But he came steadily lower and lower. Though hit and badly wounded, dropped t

ZOOM

NLA Australian Newspapers – article display

http://newspapers.nla.gov.au/ndp/del/article/11817381?searchTerm=romance+flyi

Google

Apple Yahoo! Google Maps YouTube Wikipedia News (85) Popular

Newspaper Article | NLA Australian Newspapers – arti...

AUSTRALIAN NEWSPAPERS

Home | About Us | Browse | Help | Feedback | Login / Signup

Search results | Advanced Search

romance of flying no 56 **Search Articles**

The Argus (Melbourne, Vic. : 1848-1954) (about) ◂ Saturday 10 June 1944 ▸ ◂ Page 5 S of 32 ▸

Print | Save as PDF | Save as Image

Cite: http://nla.gov.au/nla.news-article11817381

Tags (Keywords) Add New Tags

Comments Add New Comment

No comments yet.

ELECTRONICALLY TRANSLATED TEXT Fix this Text

Why may this text have mistakes? How to correct this text?

No corrections yet

IthîROMCE

iof FLLi IP G

i_

IN their early planes the Wrights obtained lateral control in flight r

by the use of warping wings. This was obviously only a makeshift j arrangement, and a new plan had to be devised. Glenn Curtiss, whose x association with **flying** began in 1903, when he began building engines * for airships, built an aeroplane in 1908 with small movable planes placed between the upper and lower wings (see sketch, upper right), which could

be tilted like elevators. They enabled the pilot to keep the machine > steady in uneven air currents and to bank when turning. Soon this plan

was incorporated with that of the Wrights. Movable flaps were placed at ' the rear edge of each wing as ailerons, an essential feature of modern I aircraft. Curtiss's 1908 biplane won the Gordon Bennett Cup at the ! first air meeting at Rheims (Prance) in 1909. It was powered by a 50

horse-power eight-cylinder engine. !

Cawthorne, MM and Bar, AIF

View entire page

IN their early planes the Wrights obtained lateral con by the use of warping wings. This was obviously onl arrangement, and a new plan had to be devised. Glenn association with flying began in 1903, when he began bui for airships, built an aeroplane in 1908 with small movable between the upper and lower wings (see sketch, upper right be tilted like elevators. They enabled the pilot to keep steady in uneven air currents and to bank when turning. was incorporated with that of the Wrights. Movable flaps the rear edge of each wing as ailerons, an essential featu aircraft. Curtiss's 1908 bipl first air meeting at Rheims (

ZOOM

NLA Australian Newspapers – article display

http://newspapers.nla.gov.au/ndp/del/search?searchTerm=romance+of+flying+no

Google

Apple Yahoo! Google Maps YouTube Wikipedia News (85) Popular

Newspaper Article | NLA Australian Newspapers – arti...

ELECTRONICALLY TRANSLATED TEXT

Need help? Keyboard Shortcuts

Save Exit Cancel

Undo Line Insert Symbol

IthîROMCE

iof FLLi IP G

i_

IN their early planes the Wrights obtained lateral control in flight r

by the use of warping wings. This was obviously only a makeshift j

arrangement, and a new plan had to be devised. Glenn Curtiss, whose x

association with flying began in 1903, when he began building engines

for airships, built an aeroplane in 1908 with small movable planes placed

between the upper and lower wings (see sketch, upper right), which

be tilted like elevators. They enabled the pilot to keep the machine >

steady in uneven air currents and to bank when turning. Soon this plan

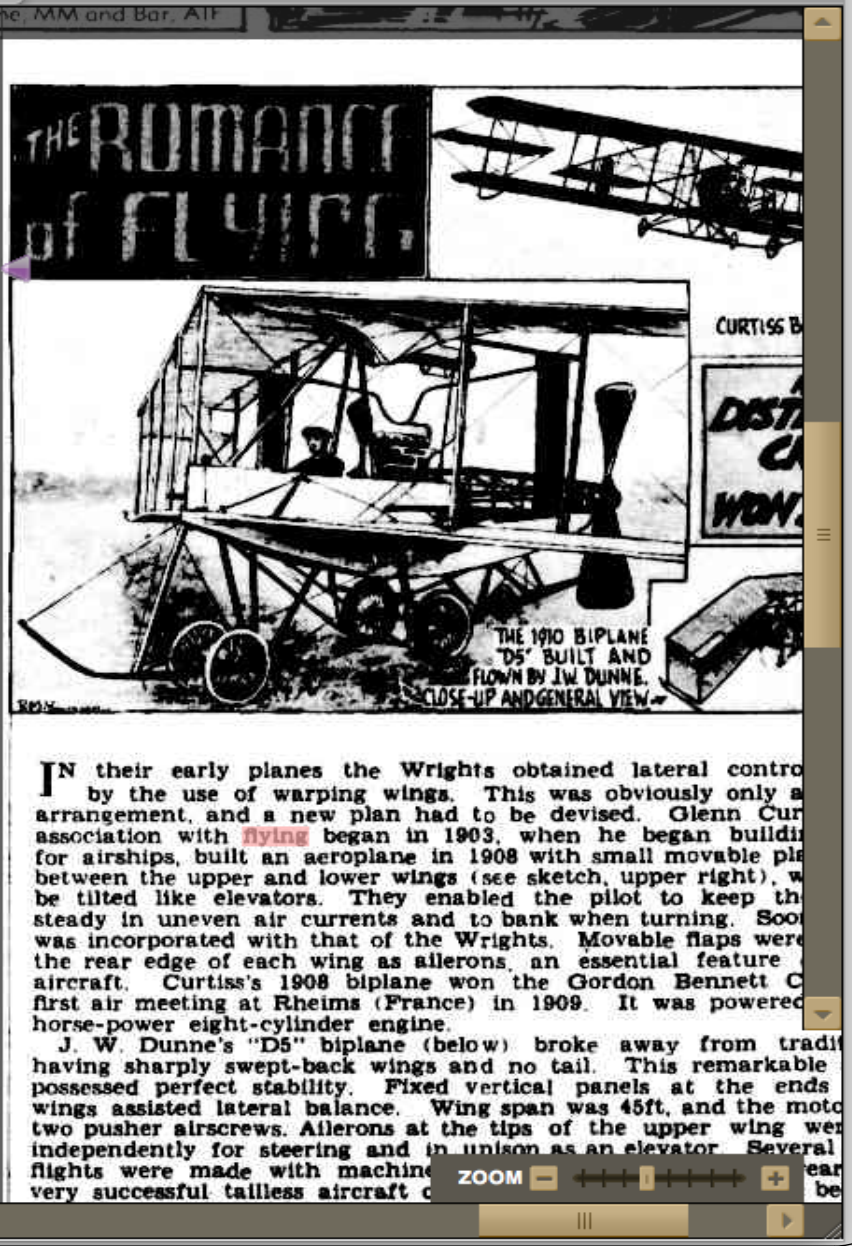
was incorporated with that of the Wrights. Movable flaps were placed at '

the rear edge of each wing as ailerons, an essential feature of modern I

aircraft. Curtiss's 1908 biplane won the Gordon Bennett Cup at the !

first air meeting at Rheims (Prance) in 1909. It was powered by a 50

horse-power eight-cylinder engine. !



NLA Australian Newspapers – article display

http://newspapers.nla.gov.au/ndp/del/article/11817381?searchTerm=romance+flyi

Google

Apple Yahoo! Google Maps YouTube Wikipedia News (86) Popular

Newspaper Article | NLA Australian Newspapers – arti...

ELECTRONICALLY TRANSLATED TEXT

Need help? Keyboard Shortcuts

Save | Save & Exit | Cancel

Undo Line | Insert Symbol

The Romance

of Flying

IN their early planes the Wrights obtained lateral control in flight

by the use of warping wings. This was obviously only a makeshift

arrangement, and a new plan had to be devised. Glenn Curtiss, whose

association with flying began in 1903, when he began building engines

for airships, built an aeroplane in 1908 with small movable planes placed

between the upper and lower wings (see sketch, upper right), which could

be tilted like elevators. They enabled the pilot to keep the machine >

steady in uneven air currents and to bank when turning. Soon this plan

was incorporated with that of the Wrights. Movable flaps were placed at '

the rear edge of each wing as ailerons, an essential feature of modern I

aircraft. Curtiss's 1908 biplane won the Gordon Bennett Cup at the !

first air meeting at Rheims (Prance) in 1909. It was powered by a 50

horse-power eight-cylinder engine. !

IN their early planes the Wrights obtained lateral control by the use of warping wings. This was obviously only a r arrangement, and a new plan had to be devised. Glenn Curtis association with flying began in 1903, when he began building for airships, built an aeroplane in 1908 with small movable plan between the upper and lower wings (see sketch, upper right), wh be tilted like elevators. They enabled the pilot to keep the steady in uneven air currents and to bank when turning. Soon was incorporated with that of the Wrights. Movable flaps were the rear edge of each wing as ailerons, an essential feature of aircraft. Curtiss's 1908 biplane won the Gordon Bennett Cu first air meeting at Rheims (France) in 1909. It was powered horse-power eight-cylinder engine.

J. W. Dunne's "D5" biplane (below) broke away from trad having sharply swept-back wings and no tail. This remarkable possessed perfect stability. Fixed vertical panels at the ends wings assisted lateral balance. Wing span was 45ft, and the mo two pusher airscrews. Ailerons at the tips of the upper wing w independently for steering and in unison as an elevator. Severa flights were made with machines of this type. In recent ye very successful tailless aircraft on somewhat similar lines have b duced.—(Contd).

rger! (GOOD-OH, PAL!

ake a hit in Austral

ver, you find the spirit of co-operation

for the common good." That's why i

s and Americans. "*Have a Coke*," says th

ows he has a comrade. From Adelaide

-Cola stands for *the p*

ZOOM

Text Correction Help

Why may this text have mistakes?

How to correct this text?

Keyboard shortcuts

To speed up the editing process, you can use keyboard shortcuts to navigate through the text and save your changes. Keyboard behaviour depends on your browser but in general:

- TAB allows you to move to the next line while Shift-TAB allows you to move to the previous line.
- Enter allows you to move to the next line.
- Up and Down arrow keys also allow you to move from line to line.
- Left and Right arrows allow you to move within a line or a text area. Use the Home and End keys to position the cursor at the beginning and end of the input field respectively.
- F12 allows you to move from word to word.
- Windows: Alt + s focuses the Save button, Alt + e focuses the Exit/Save & Exit button and Alt + x focuses the Cancel button. Hit Enter to perform the action. (These rely on browser "access key" functionality and may not work with all browsers.)
- MacOS X: Ctrl + s is equivalent to Save, Ctrl + e to Exit/Save & Exit and Ctrl + x to Cancel.

Special Characters

Missing Text

Multi-Page Articles

Searching on Your Changes

Close

ELECTRONICALLY TRANSLATED TEXT Fix this Text

Why may this text have mistakes? How to correct this text?

1 correction by anonymous - Show corrections

The **Romance**

of **Flying**

IN their early planes the Wrights obtained lateral control in flight

by the use of warping wings. This was obviously only a makeshift arrangement, and a new plan had to be devised. Glenn Curtiss, whose association with **flying** began in 1903, when he began building engines for airships, built an aeroplane in 1908 with small movable planes placed between the upper and lower wings (see sketch, upper right), which could

be tilted like elevators. They enabled the pilot to keep the machine > steady in uneven air currents and to bank when turning. Soon this plan

was incorporated with that of the Wrights. Movable flaps were placed at ' the rear edge of each wing as ailerons, an essential feature of modern I aircraft. Curtiss's 1908 biplane won the Gordon Bennett Cup at the ! first air meeting at Rheims (Prance) in 1909. It was powered by a 50

horse-power eight-cylinder engine. !

IN their early planes the Wrights ob
by the use of warping wings. This was obviously only a
arrangement, and a new plan had to be devised. Glenn Curti
association with flying began in 1903, when he began building
for airships, built an aeroplane in 1908 with small movable plan
between the upper and lower wings (see sketch, upper right), wh
be tilted like elevators. They enabled the pilot to keep the
steady in uneven air currents and to bank when turning. Soon
was incorporated with that of the Wrights. Movable flaps were
the rear edge of each wing as ailerons, an essential feature of
aircraft. Curtiss's 1908 biplane won the Gordon Bennett Cu
first air meeting at Rheims (France) in 1909. It was powered
horse-power eight-cylinder engine.

J. W. Dunne's "D5" biplane (below) broke away from trad
having sharply swept-back wings and no tail. This remarkable
possessed perfect stability. Fixed vertical panels at the ends
wings assisted lateral balance. Wing span was 45ft, and the mo
two pusher airscrews. Ailerons at the tips of the upper wing w
independently for steering and in unison as an elevator. Severa
flights were made with machines of this type. In recent yea
very successful tailless aircraft on somewhat similar lines have b
duced.—(Contd).

Urls

Harvester and Newspapers Service source code available from National Library of Australia's Code repository

https://code.nla.gov.au/

Newspapers Application

http://newspapers.nla.gov.au/ndp/del/home

http://trove.nla.gov.au/

Newspapers project details

http://www.nla.gov.au/ndp/project_details/