

Manual

Page 1 of 6



WEB EXTRACTOR

Manual



TABLE OF CONTENTS

1	AP	P documentation	3
_	1.1	HOW IT WORKS	
	1.2	Input data	4
	1.3	Output data	4
	1.4	Basic workflow example	5
2	AP	I documentation	6



WEB EXTRACTOR

1 APP DOCUMENTATION

1.1 HOW IT WORKS

Web Extractor APP is an acquisition task that enables to extract information like product records, news, comments and other kind of information from web sites like e-commerce, online media and other web sites.

It requires as input one or more wep files properly created with the Web Extractor Modeler, an ALTILIA tool that enables to records action and rules for exploring a web site and its contents (see the Web Extractor Modeler User Guide to know more about this topic).



WEB EXTRACTOR

Manual

Page 4 of 6

1.2 INPUT DATA

This APP allows You:

Web Extractor Wep files

Parallelism degree 1 -

- to select one or more wep file that explain to MANTRA how and where capture data from a web site;
- ✓ specify the parallelism degree.



Step 1 – click in the wep files field



Step 3 – selecting wep files



		1.00	Web Extractor		
	Upload window		X		
	Warning: wep files only admitted.	Succes	amadori, weg		
	(*) Upload			1970000006606	22
	(A				
	Dron Files Here				
0	L				
0-	Parameters				
	guestin22c1e08e-M75-40a5-b61a-4/11b4bec842/weplanadori.wep				

Step 4 – wep file selected

Figure 1 – how to select input data.

Once selected the wep files, you must select the parallelism degree.

1.3 OUTPUT DATA

This APP gives back a set of variable elements; they depends on the wep files you are using.

Then isn't possible describe the output data in this user guide.

But as example, if you made a wep file that capture data about products of a web site, you can have as output fields having information on product name or description or price and so one.

Those output fields will be available for the APPs that follow the Web Extractor APP.

These APPs can be Sentiment-Extractor APP for Brand Reputation Analysis, or MANTRA Language APP for transforming semi-structured data coming from the Web Extractor APP in structured data on which to make subsequent analysis like price comparison (Price Intelligence Analysis as example).



1.4 BASIC WORKFLOW EXAMPLE

Web Extractor needs some transformation and / or normalization APP after in the workflow that can process information captured, so you can create a flow as shown in the following figure left below.

In this example we have a market intelligence use case example, where APPs, starting from semi structured data, can compare product price from different web sites.

You can set input parameter and obtain the results by the to excel APP as shown in the following figure at the right side:



Figure 2 - web extractor workflow example.

A	8	C D	E	F		ö	
1 marcalpr	ice CodVan	CodArt EAN	prezzoComp	otito BATETIME RILEVAZIONE	e utiliprico		urlCompetitor
2 FUJFILM	16084967	100415680 45474101	31086 249.90	2013-09-12 05:53:05	http://www.aprice.it/fotocamara-	digitali-FUJIFILM/d-2965156	http://www.menclick.it/schede/tujif :
3 FUUPEM	16218883	f101251004 25474102	00119 75.88	2013-03-12 05:53:05	http://www.eprice.it/fotocamere-	digitali-PUUIPILM/6-6070804	http://www.msnclick.it/schede/tujf_
4 FILIFILM	16217485	101251052 35474102	00157 75 86	2013-03-12 05 53 05	http://www.eprice.it/fistucienses-	dipitali-FiLUFILMAI-5070789	http://www.comclick.it/schede/kpl
5 FUUFILM	2004343	100902000 45474102	02030 217 02	2013-03-12 03:42:07	http://www.eprice.it/fotocamere-	digitali-FUJITUMid-4291205	http://www.monclick.it/schede/tujf
E FUJFILM	96262648	101251005 45474102	17063 75.88	2013-03-12 05:43:53	http://www.eprice.it/btocamere-	dicitali FUJIFILM/d-5070785	http://www.manclick.it/scheda/tuff
/ FUIFILM	162/3866	1010/6696 264/4102	22624 868.28	201		Itali HUJHUMid 4622/09	http://www.menslick.it/schode/tupt
8 FUJFILM	16283381	101161657 45474103	29370 56.79	201 The Exce	Report shows	Itall-FUJIFILM/d-1798911	http://www.menclick.it/schede/tuif
5 FUJFILM	16284082	101161941 25474102	29417 85.79	201		itali-FUJPEM/d4798320	http://www.monclick.it/schede/hjif
10 FILIFIM	96786553	101217978 45474102	32923 57 45	201 Structu	ired product	itali-FLLIFILM61-4902185	http://www.come.fr.h.it/schede/kgif
11 FUUFLM	96207234	101218505 35474102	33001 07.45	201 instance	ac with prices	itali-FUJIFILMid-4903431	http://www.monclick.it/schede/fuif
12 FUJFUM	96307204	101218509 45474102	30525 109.62	28 HISLANC	es with prices	itali FUJIFILM/d-4983458	http://www.menclick.it/schede/tust
TS FUIFILM	"16.JU/388	101218314 464/4102	91007 109.62	2012 201 202 20 20 20 20 20 20 20 20 20 20 20 20		usitali FGJHLMid 498341/	http://www.menclick.d/schedo/tust
14 FUJFLM	XP60BL	101192520 45474103	MD412 161.78	2013-09-12 05:49:53	http://www.aprice.it/btocamara-	digitali-FUJIFILM/d-1983508	http://www.menclick.it/schede/tuif
15 FURFILM	XP80BL	101192520 25474102	40412 151.78	2013-03-12 05:43:53	http://www.eprice.it/fotocamere-	dgtall/FUJPEMid-4983508	http://www.menclick.it/schede/tojif
15 FILIFEM	XPSDYF	101192111 45474102	40564 151 78	2013-09-12 05 49 53	http://www.eprice.ildfalationere-	digitali-FULBFILM61-4983484	http://www.mustick.d/schede/hgf
17 CANON	50450009	100572312 07145745	65835 276.02	2013-03-12 00:43:40	http://www.eprice.it/fotocamere-	digitali-CANON/d-3252514	http://www.monclick.it/schede/can
18 CANON	52498009	100985941 87145745	77739 552 46	2013-03-12 05:43:43	http://www.sprice.it/fotocamere-	dicitali CANDN/5-4758440	http://www.monclick.8/scheds/can
19 CANON	69009011	100946992 8/146746	30432 188.61	2013-09-12-05:44:23	http://www.eprice.it/fotecomera-	6 gitali C/MUN/6 4226821	http://www.menclick.it/schedo/can
20 CANON	60258006	100544872 87146745	80937 246.64	2013-09-12 05:49:40	http://www.aprice.it/fotocamara-	digitali-CANON/5-4172241	http://www.menclick.it/schede/can
21 CANON	60288006	100544595 87145745	81019 246.64	2013-03-12 05:43:43	http://www.eprice.it/fotocamere-	dgitali-CANON/54172195	http://www.monclick.it/schede/carr
77 CANON	60315006	100544873 87145745	81057 246.64	2013-09-12 09 49 43	http://www.eprice.it/futucanera-	digitali-CANON/3-4172203	http://www.munclick.it/schedelcare
23 CANON	01450011	100540401 07145745	00753 324.37	2013-03-12 03:43:40	http://www.eprice.it/fotocamere-	digitali-CANON/3-4172256	http://www.monclick.it/schede/can
24 CANON	61508011	100548745 87145745	78507 143.15	2013-03-12 05:42:07	http://www.sprice.it/lotocamere-	dicitali CANON/64204736	http://www.menclick.it/schede/can
25 CAWON	61945099	100946994 8/146/46	30470 188.61	2013 09 12 05 49 25	http://www.eprice.it/totecamore-	digitali C/MON/d 4229822	http://www.menclick.it/schede/can
26 CANON	61958011	100545995 87145745	80517 176.88	2013-09-12 05:41:53	http://www.aprice.it/fotocamara-	digitali-CANON/d-4174670	http://www.menclick.h/schede/can
27 CANON	61968011	100546054 87145745	80655 188.61	2013-03-12 05:44 23	http://www.eprice.it/fotocamere-	digitali-CANON/54174702	http://www.monclick.it/schede/can
23 CANON	63539009	101057239 37145745	85311 180.13	2013-03-12 05 49 25	http://www.eprice.it/foliocentere-	digitali-CANONAL4609566	http://www.marscick.it.d/schede/s.am
29 CANON	63543087	101057240 07145749	0G370 99.70	2013-03-12 03:43:23	http://www.eprice.it/fotocamere-	digitali-CANON/3-4503577	http://www.monclick.it/schede/can
20.010480744	E060GID4C	Penerspean Street of	ANALS OF RC	1012031202120	have the owners and it distances and	And all PARTIALLY PRESS.	hits floore manufact divide dates

Figure 3 - the excel format of results.



WEB EXTRACTOR

Manual

Page 6 of 6

2 API DOCUMENTATION

For information about how to use Article Extractor API in your application, send us a message to info@altiliagroup.com.