



WEB EXTRACTOR

Manual

TABLE OF CONTENTS

1 APP documentation.....	3
1.1 HOW IT WORKS.....	3
1.2 Input data	4
1.3 Output data	4
1.4 Basic workflow example.....	5
2 API documentation.....	6

1 APP DOCUMENTATION

1.1 HOW IT WORKS

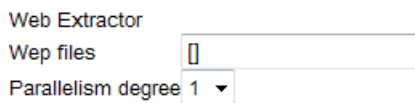
Web Extractor APP is an acquisition task that enables to extract information like product records, news, comments and other kind of information from web sites like e-commerce, online media and other web sites.

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).

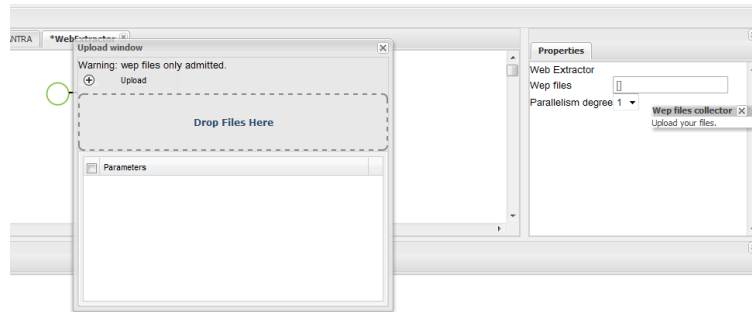
1.2 INPUT DATA

This APP allows You:

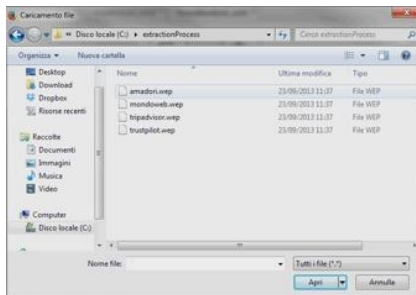
- ✓ to select one or more wep file that explain to MANTRA how and where capture data from a web site;
- ✓ specify the parallelism degree.



Step 1 – click in the wep files field



Step 2 – upload windows opening



Step 3 – selecting wep files



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:

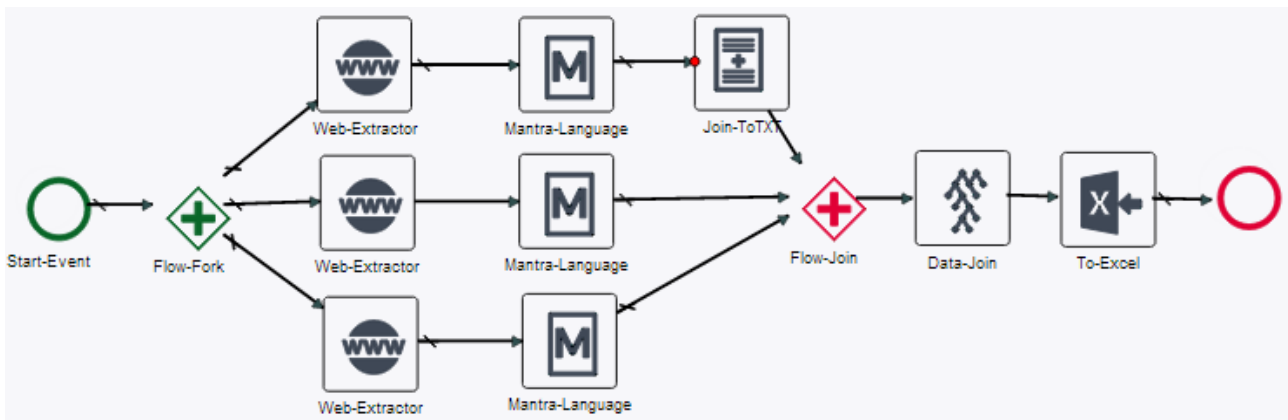


Figure 2 - web extractor workflow example.

A	B	C	D	E	F	G
marca	price	CodVen	CodArt	EAN	prezzoCompetitor	URL
1	marca	price	CodVen	CodArt	EAN	prezzoCompetitor
2	FUJIFILM	96884967	900415690	547410131086	249.50	2013-09-12 05:52:05
3	FUJIFILM	96218883	900251004	547410200119	75.88	2013-09-12 05:53:05
4	FUJIFILM	96217685	900251059	547410200157	75.88	2013-09-12 05:53:05
5	FUJIFILM	9084343	900502000	547410202030	217.02	2013-09-12 05:42:07
6	FUJIFILM	96252548	900251905	547410217063	75.88	2013-09-12 05:43:53
7	FUJIFILM	96217686	900161698	547410222824	89.28	201
8	FUJIFILM	96283381	900161657	547410229370	55.75	201
9	FUJIFILM	96284062	900161941	547410229417	55.75	201
10	FUJIFILM	96289553	900161797	547410230929	57.45	201
11	FUJIFILM	96207234	900218020	547410230001	57.45	201
12	FUJIFILM	96207234	900218020	547410230025	102.62	20
13	FUJIFILM	96207234	900218020	547410230025	102.62	20
14	FUJIFILM	XP50BL	900119250	547410240412	151.78	2013-09-12 05:43:53
15	FUJIFILM	XP80BL	900119250	547410240412	151.78	2013-09-12 05:43:53
16	FUJIFILM	XP50YF	900119211	547410240504	151.78	2013-09-12 05:43:53
17	CANON	50450089	900572312	8714574525035	276.02	2013-09-12 05:43:43
18	CANON	52495099	900585941	8714574577739	552.46	2013-09-12 05:43:43
19	CANON	59909011	900544594	8714574480432	188.81	2013-09-12 05:43:23
20	CANON	60285096	900544872	8714574580837	246.64	2013-09-12 05:43:43
21	CANON	60289096	900544594	8714574481019	248.64	2013-09-12 05:43:43
22	CANON	60315086	900544872	8714574581057	246.64	2013-09-12 05:43:43
23	CANON	61402011	900544401	8714574500753	354.37	2013-09-12 05:43:43
24	CANON	61505011	900544874	8714574578507	342.35	2013-09-12 05:42:07
25	CANON	61945099	900544594	8714574481049	188.81	2013-09-12 05:43:23
26	CANON	61958011	900545995	8714574580617	176.88	2013-09-12 05:41:53
27	CANON	61965011	900546054	8714574580955	188.81	2013-09-12 05:44:23
28	CANON	63535089	900557739	8714574583111	180.15	2013-09-12 05:43:23
29	CANON	63540087	900557240	8714574583373	95.70	2013-09-12 05:43:23

Figure 3 - the excel format of results.

2 API DOCUMENTATION

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