

# “*Obtorto\_collo*-Data da ARPA Lombardia”

Andrea Trentini - <http://atrent.it>  
Dipartimento di Informatica - UniMi





# Livelli (citt.dig. & tecnociv.)





... Il alcune stazioni della rete di Qualità dell'Aria, pur normalmente rilevati, non risultano pubblicati sul sito per attività di manutenzione ...

Home > Aria > Rete di rilevamento

Chi siamo

Settori

Dipartimenti

Servizi

LIRP

Ufficio stampa

Bandi, concorsi

Area download

Contatti



Oltre la rete ...

- » Inquinanti nel bacino padano in Italia ed in Europa
- » PM10 a confronto
- » Il sito della Qualità dell'Aria della Regione

I dati storici ...

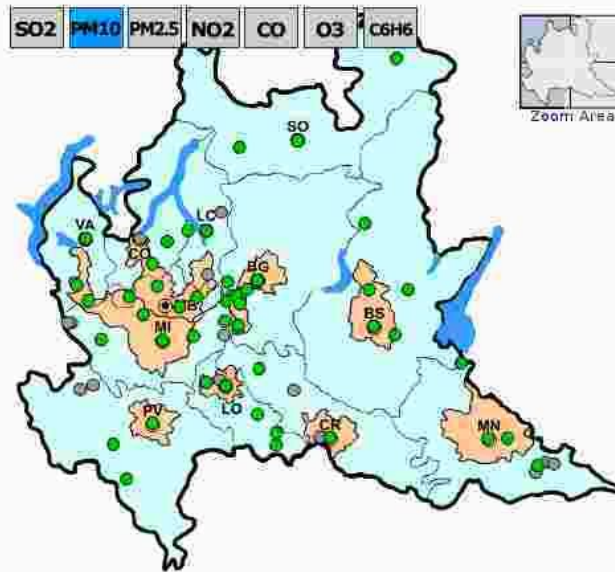
- » L'evoluzione negli anni
- » Il trend del PM10
- » Il confronto con i limiti della normativa
- » Download archivio storico

Meteo ...

- » L'importanza delle condizioni meteorologiche
- » Meteo & Inquinanti: previsioni per i prossimi giorni

Rete di rilevamento ...

Situazione relativa a Lunedì 10 Febbraio 2014



L'aria nelle stazioni ...

Zona del territorio / Provincia

Città

Centralina

VAI

PM10, la media delle concentrazioni negli agglomerati urbani.

Mappe di qualità aria ...

- » Il sistema modellistico utilizzato
- » Mappe di concentrazione giornaliere e valori comunali stimati
- » L'inventario delle emissioni INEMAR

SCARICA!

# Tim Berners-Lee



E allora...  
“webstacles”!  
(web+obstacles)

<http://www.w3.org/DesignIssues/LinkedData.html>

Qualità dell'Aria

ARPA LOMBARDIA  
Agenzia Regionale per la Protezione dell'Ambiente

Regione Lombardia

SITO DI QUALITÀ DELL'ARIA di ARPA Lombardia

Home > Aria > Richiesta dati

Archivio dei dati rilevati

Dati storici e realtime

Provincia:

Città:

Centralina:

Inquinante:

Sensori:

Seleziona:

Dal:  Al:

E-Mail:

Controlla:

Richiedi

L'aria nelle stazioni

Zona del territorio / Provincia:

Città:

Centralina:

VAT

PM10, la media delle concentrazioni negli agglomerati urbani

I dati storici

- L'evoluzione negli anni
- Il trend del PM10
- Il confronto con i limiti della normativa
- Download archivio storico

Oltre la rete

- Inquinanti nel bacino padano in Italia ed in Europa
- PM10 a confronto
- Il sito della Qualità dell'Aria della Regione

Meteo

- L'importanza delle condizioni meteorologiche
- Meteo & Inquinanti: previsioni per i prossimi giorni

Documentazione

- Relazioni annuali provinciali
- Campagne mezzi mobili
- La valutazione modellistica annuale

~150 staz  
X  
~5 item  
X  
~30 anni


=====

**>20k  
query  
(mail)**

# Soluzione?


- PASSATO
  - Nessuna!
    - Salvo avere soldi (captcha human solving)
- FUTURO (che comincia dal 2005!)
  - Scraping + Scripting





ARPA LOMBARDIA  
Agenzia Regionale per la Protezione dell'Ambiente

**Qualità dell'Aria**



Regione Lombardia

... non risultano pubblicati sul sito per attività di manutenzione straordinaria sui sistemi di trasmissione ;

» Home » Aria » Agglomerato di Milano

Chi siamo

Settori

Dipartimenti

Servizi

LIRP


Ufficio stampa

Bandi, concorsi

Area download

Contatti

... Zona del territorio / Provincia ...



Dati del 10 Febbraio 2014

**Agglomerato di Milano**

... L'aria nelle stazioni ...

Zona del territorio / Provincia  
Agglomerato di Milano

Città


Centralina

VAI

» PM10, la media delle concentrazioni negli agglomerati urbani.

... Gli inquinanti monitorati ...

Inquinante Rilevamento Soglie / Limiti Unità di misura	<b>SO<sub>2</sub></b>	<b>PM10</b>	<b>PM2.5</b>	<b>NO<sub>2</sub></b>	<b>CO</b>	<b>O<sub>3</sub></b>	<b>Benzene</b>
	Media giornaliera	Media giornaliera	Media giornaliera	Massimo giornaliero	Max media mobile 8h giornaliera	Massimo giornaliero	Media giornaliera
	val. limite 125	val. limite 50		val. limite 200	val. limite 10	soglia di informaz. 180 soglia di allarme 240	
Comune Centralina	µg/m3	µg/m3	µg/m3	µg/m3	mg/m3	µg/m3	µg/m3
<b>Milano</b>							
<b>Arese</b>	--	17	--	64	--	48	--
<b>Cinisello Balsamo</b>	--	--	--	nd	--	--	--
<b>Cormano</b>	nd	--	--	86	--	35	--
<b>Corsico</b>	--	--	--	101	1.1	40	--
<b>Garbagnate</b>	--	--	--	93	--	--	--



# Scraping...

- wget is your friend...

```
/usr/bin/wget -q
```

```
-O $DATA/$(date +%Y%m%d%H%M)_aria$num.html
```

```
http://ita.arpalombardia.it/ITA/qaria/lista$num.asp 2>&1
```

- ... and cron is your ally

```
0 1 * * * /root/aria/aria.sh
```

(since 2005)

# RAW DATA

```
-rw-r--r-- 1 root root 8.3K Feb 11 01:00 201402110100_aria01.html.gz
-rw-r--r-- 1 root root 6.7K Feb 11 01:00 201402110100_aria02.html.gz
-rw-r--r-- 1 root root 6.1K Feb 11 01:00 201402110100_aria03.html.gz
-rw-r--r-- 1 root root 5.8K Feb 11 01:00 201402110100_aria04.html.gz
-rw-r--r-- 1 root root 5.5K Feb 11 01:00 201402110100_aria05.html.gz
-rw-r--r-- 1 root root 5.7K Feb 11 01:00 201402110100_aria06.html.gz
-rw-r--r-- 1 root root 6.2K Feb 11 01:00 201402110100_aria07.html.gz
-rw-r--r-- 1 root root 5.6K Feb 11 01:00 201402110100_aria08.html.gz
-rw-r--r-- 1 root root 5.4K Feb 11 01:00 201402110100_aria09.html.gz
-rw-r--r-- 1 root root 5.5K Feb 11 01:00 201402110100_aria10.html.gz
```

# (sorgente)

```
...class="linktesto"><B>Cormano</B><br><font color="#666666"></font></a></td>
<td width="1" bgcolor="#003366"></td>
<td width="54" align="right" class="testobox">nd</td>
<td width="18" valign="middle" align="center"></td>
<td width="1" bgcolor="#003366"></td>
<td width="54" align="right" class="testobox">--</td>
<td width="18" valign="middle" align="center"></td>
<td width="1" bgcolor="#003366"></td>
<td width="54" align="right" class="testobox">--</td>
<td width="18" valign="middle" align="center"></td>
<td width="1" bgcolor="#003366"></td>
<td width="54" align="right" class="testobox">86</td>
<td width="18" valign="middle" align="center"></td>...
```

Estrai!

# ScrApting

- BASH and Unix FILTERS!!!
  - (I'm an old dog, sorry)
    - Today: use Scrapy et similia

```
... | sed 's/<br>/ /g' | vilistextum -a - - | tr -d '\177-\377' |  
sed 's/</(min.di)/g' | tee tmp.txt | tr -s "\n" " " | sed -f  
sed.script | grep -ivf ../righePost.txt | tr -s " " ";" | sort >  
$FILE.dati
```

...

# CSV Data

```
-rw----- 1 atrentini atrentini 84157 Dec 16 11:51 Milano_Abbiategrasso.csv
-rw----- 1 atrentini atrentini 27428 Dec 16 11:51 Milano_Juvara.csv
-rw----- 1 atrentini atrentini 91904 Dec 16 11:51 Milano_Liguria.csv
-rw----- 1 atrentini atrentini 94659 Dec 16 11:51 Milano_Marche.csv
-rw----- 1 atrentini atrentini 1871 Dec 16 11:51 Milano_Messina.csv
-rw----- 1 atrentini atrentini 88139 Dec 16 11:51 Milano_Parco_Lambro.csv
-rw----- 1 atrentini atrentini 84731 Dec 16 11:51 Milano_Pascal_Citt_Studi.csv
-rw----- 1 atrentini atrentini 111540 Dec 16 11:51 Milano_Senato.csv
-rw----- 1 atrentini atrentini 95907 Dec 16 11:51 Milano_Verziere.csv
-rw----- 1 atrentini atrentini 109921 Dec 16 11:51 Milano_Zavattari.csv
```

# CSV Data (detail)

Milano Zavattari.csv

Data,SO2,PM10,PM2.5,NO2,CO,O3,Benzene

2005/12/5, , ,143,3.2, ,6.8,

2005/12/6, , ,196,4.6, ,9.0,

2005/12/7, , ,135,4.5, ,8.2,

2005/12/8, , ,165,5.4, ,10.6,

2005/12/9, , ,190,4.5, ,8.0,

2005/12/10, , ,150,4.7, ,7.1,

2005/12/11, , ,135,3.8, ,6.3,

2005/12/12, , ,155,3.5, ,6.2,

2005/12/13, , ,103,3.0, ,5.4,

...

...

2013/12/5, , , ,223,3.5, ,(min.di)3,

2013/12/6, , , ,167,3.2, ,(min.di)3,

2013/12/7, , , ,121,1.7, ,(min.di)3,

2013/12/8, , , ,112,1.9, ,(min.di)3,

2013/12/9, , , ,87,2.1, ,3.5,

2013/12/10, , , ,147,3.2, ,6.3,

2013/12/11, , , ,190,4.6, ,6.7,

2013/12/12, , , ,133,4.6, ,6.9,

2013/12/13, , , ,139,2.8, ,4.9,

2013/12/14, , , ,143,3.0, ,4.5,

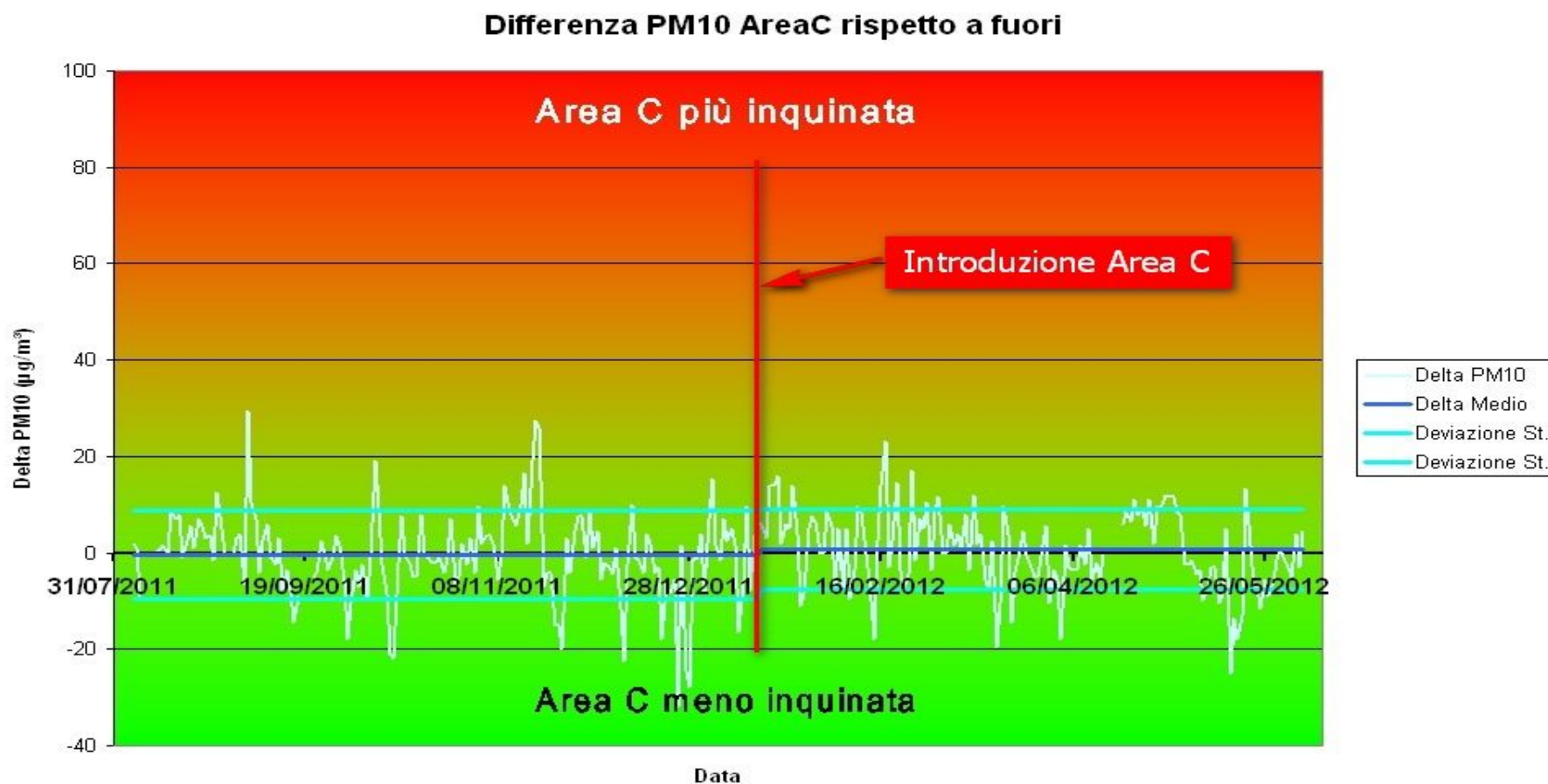
# Pregi e difetti

- Verifica storica dei dati!!!
- Gira dal 2005
- Grafici generati in automatico (gnuplot) con i dati aggiornati
- Procedimento migliorabile?
  - Il già citato Scrapy (ma è in Python...)
- Storage dati grezzi non ottimizzato



# Risultati ottenuti

Esempio: “effetti di AreaC” = 0!



# Grazie!

Dati e grafici qui:

<http://arcipelagoareac.it>

<http://motocivismo.it>

Andrea Trentini - <http://atrent.it>

Dipartimento di Informatica - UniMi