

Tecnología de Desarrollo de Software II

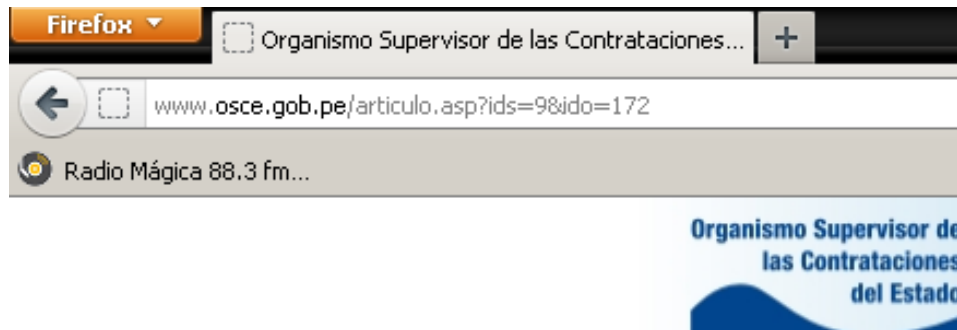
Descarga de Resoluciones Emitidas por el Tribunal del Organismo Supervisor de las Contrataciones del Estado

<http://www.osce.gob.pe>

Integrantes:

Muga Ampuero, Luis
Salinas Francia, Antonio

1. Ingresamos a la dirección **<http://www.osce.gob.pe/articulo.asp?ids=9&ido=172>** desde nuestro navegador Firefox.



2. Presionamos **Ctrl+U** y se nos mostrará el código fuente de la página.

```
1 <html>
2 <head>
3   <meta http-equiv="Content-Type" content="text/html; charset=iso-8859-1" />
4   <!-- Inicio Titulo de Pagina-->
5   <title>Organismo Supervisor de las Contrataciones del Estado</title>
6   <!-- Fin Titulo de Pagina-->
7   <!-- <title>Consejo Superior de Contrataciones y Adquisiciones del Estado</title-->
8   <link href="includos/css/portat consuode.css" rel="stylesheet" type="text/css" />
9   <script type="text/javascript" src="includos/menu/anylink.js"></script>
10  <script src="includos/js/ac runactivecontent.js" type="text/javascript"></script>
11  <script type="text/javascript">
12    <!--
13    function MM_preloadImages() { //v3.0
14      var d=document; if(d.images){ if(!d.MM_p) d.MM_p=new Array();
15        var i,j=d.MM_p.length,a=MM_preloadImages.arguments; for(i=0; i<a.length; i++)
16          if (a[i].indexOf("#")!=0){ d.MM_p[j]=new Image; d.MM_p[j++].src=a[i];}}
17      }
18      function MM_swapImgRestore() { //v3.0
19        var i,x,a=document.MM_sr; for(i=0;a&&i<a.length&&(x=a[i])&&x.oSrc;i++) x.src=x.oSrc;
20      }
21      function MM_findObj(n, d) { //v4.01
22        var p,i,x; if(!d) d=document; if((p=n.indexOf("?"))>0&&parent.frames.length) {
23          d=parent.frames[n.substring(p+1)].document; n=n.substring(0,p);
24        } if(!(x=d[n])&&d.all){ for (i=0; i<d.forms.length; i++) x=d.forms[i][n];
25        } if(!x){ for(i=0; i<d.layers.length; i++) x=MM_findObj(n,d.layers[i].document);
26        } if(!x && d.getElementById) x=d.getElementById(n); return x;
27      }
28      function MM_swapImage() { //v3.0
29        var i,j=0,x,a=MM_swapImage.arguments; document.MM_sr=new Array; for(i=0; i<(a.length-2); i+=3)
30          if ((x=MM_findObj(a[i]))!=null){document.MM_sr[j++]=x; if(!x.oSrc) x.oSrc=x.src; x.src=a[i+2];}
31        }
32      function fncShowHide(sh7) {
33        if (document.getElementById) {
34          obj = document.getElementById(sh7);
35          if (obj.style.display == "none") {
36            obj.style.display = "block";
37          }
38        }
39      }
40    </script>
41  </head>
42  <body>
43    <div id="contenedor">
44      <div id="encabezado">
45        <div id="logo">
46          <img alt="Logo of the Organismo Supervisor de las Contrataciones del Estado" data-bbox="625 240 790 295"/>
47        </div>
48      </div>
49    </div>
50  </body>
51 </html>
```

3. Seleccionamos todo el código (**Ctrl+A**) y lo copiamos en un archivo en un nuevo archivo de texto (mkdir /root/osce, nano pagina):

```
root@gpsapp:~/osce
[root@gpsapp ~]# mkdir /root/osce
[root@gpsapp ~]# cd /root/o
-bash: cd: /root/o: No such file or directory
[root@gpsapp ~]# cd /root/osce/
[root@gpsapp osce]# nano pagina
```

4. Pegamos todo el contenido seleccionado y cerramos el archivo (**Ctrl+X**).

5. Al revisar el código nos fijamos que los enlaces figuran como un 5to campo separado por comillas (" "), hacemos uso de awk para separarlo y redirigimos la salida a un archivo llamado "enlaces":

```
<td width="386"><a href="DescargaHit.asp?dir=userfiles/archivos/RT001-050.zip&nom=RT001-050.zip" target="_blank" class="link_artiuculo_interno">Desde Nº 001 - Nº 296</a></td>
```

Código: awk 'BEGIN {FS="\""} /DescargaHit*/ {print \$4}' pagina > enlaces

6. Quedándonos las siguientes líneas en el archivo "enlaces":

```
...
DescargaHit.asp?dir=userfiles/archivos/151-200.zip&nom=151-200.zip
DescargaHit.asp?dir=userfiles/archivos/101-150.zip&nom=101-150.zip
DescargaHit.asp?dir=userfiles/archivos/051 - 100.zip&nom=051 - 100.zip
DescargaHit.asp?dir=userfiles/archivos/RT001-050.zip&nom=RT001-050.zip
...
```

7. Agregamos el nombre de dominio a los enlaces y lo guardamos en el archivo "enlacesrevisados"; quedándonos las líneas:

Código: awk '{print "http://www.osce.gob.pe/" \$0}' enlaces > enlacesrevisados

```
http://www.osce.gob.pe/DescargaHit.asp?dir=userfiles/archivos/RESOL. (451-500) 2012.zip&nom=RESOL.
(451-500) 2012.zip
http://www.osce.gob.pe/DescargaHit.asp?dir=userfiles/archivos/RESOL. (401-450) 2012.zip&nom=RESOL.
(401-450) 2012.zip
http://www.osce.gob.pe/DescargaHit.asp?dir=userfiles/archivos/RESOL. (351-400) 2012.zip&nom=RESOL.
(351-400) 2012.zip
http://www.osce.gob.pe/DescargaHit.asp?dir=userfiles/archivos/RESOL. (301-350) 2012.zip&nom=RESOL.
(301-350) 2012.zip
http://www.osce.gob.pe/DescargaHit.asp?dir=userfiles/archivos/RESOL. (251-300) 2012.zip&nom=RESOL.
(251-300) 2012.zip
http://www.osce.gob.pe/DescargaHit.asp?dir=userfiles/archivos/RESOL. (201-250) 2012.zip&nom=RESOL.
(201-250) 2012.zip
```

8. Ahora convertimos cualquier espacio en blanco por su respectivo equivalente y hacemos uso de **wget** para descargar la lista de ficheros; borramos los archivos usados anteriormente:

Código:

awk '{gsub(" ", "%20", \$0); print \$0;}' enlacesrevisados > enlacesrevisados2

wget -i enlacesrevisados2

rm -f pagina enlaces enlacesrevisados

```
root@gpsapp:~/osce
--2012-05-13 01:59:31-- http://www.osce.gob.pe/userfiles/archivos/RESOL.%20%284
51-500%29%202012.zip
Reusing existing connection to www.osce.gob.pe:80.
HTTP request sent, awaiting response... 200 OK
Length: 1101612 (1.0M) [application/x-zip-compressed]
Saving to: âRESOL. (451-500) 2012.zipâ

100%[=====] 1,101,612 199K/s in 5.4s

2012-05-13 01:59:37 (199 KB/s) - âRESOL. (451-500) 2012.zipâ

--2012-05-13 01:59:37-- http://www.osce.gob.pe/DescargaHit.asp?dir=userfiles/ar
chivos/RESOL.%20(401-450)%202012.zip&nom=RESOL.%20(401-450)%202012.zip
Reusing existing connection to www.osce.gob.pe:80.
HTTP request sent, awaiting response... 302 Object moved
Location: userfiles/archivos/RESOL.%20%28401-450%29%202012.zip [following]
--2012-05-13 01:59:37-- http://www.osce.gob.pe/userfiles/archivos/RESOL.%20%284
01-450%29%202012.zip
Reusing existing connection to www.osce.gob.pe:80.
HTTP request sent, awaiting response... 200 OK
Length: 4715608 (4.5M) [application/x-zip-compressed]
Saving to: âRESOL. (401-450) 2012.zipâ

7%[==>] 376,535 201K/s
```

9. Podemos automatizar esto poniendo todos los pasos en un archivo bash:

Código:

```
#!/bin/bash
awk 'BEGIN {FS="\""} /DescargaHit*/ {print $4}' pagina > enlaces
awk '{print "http://www.osce.gob.pe/" $0}' enlaces > enlacesrevisados
awk '{gsub(" ", "%20", $0); print $0;}' enlacesrevisados > enlacesrevisados2
wget -i enlacesrevisados2
rm -f pagina enlaces enlacesrevisados
```