Web Scraping

@alvaro_aguirre

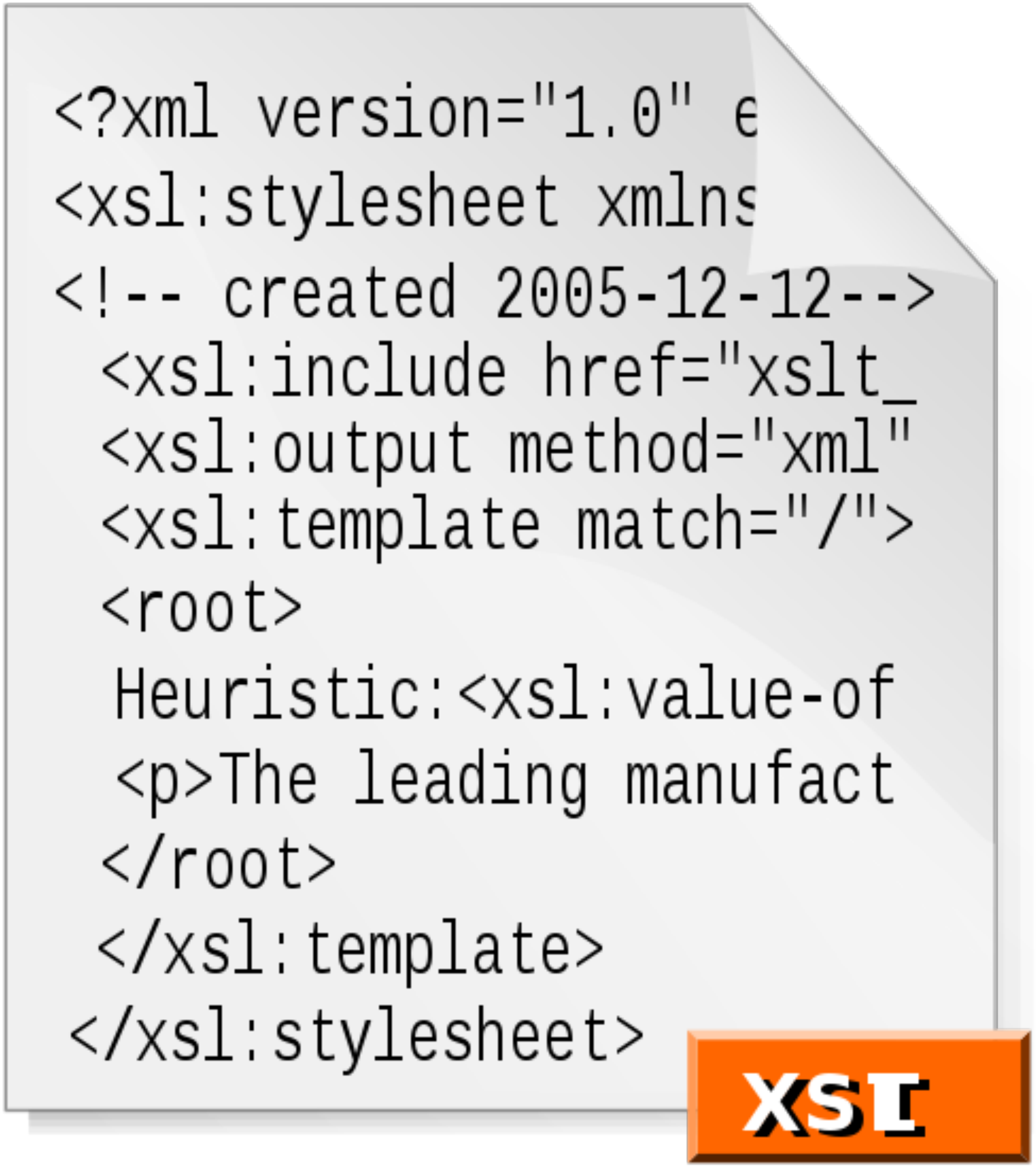
In search of our cosmic origins...

PythonChile
Comunidad de Profesionales

KUNDART

www.kundart.com

python™



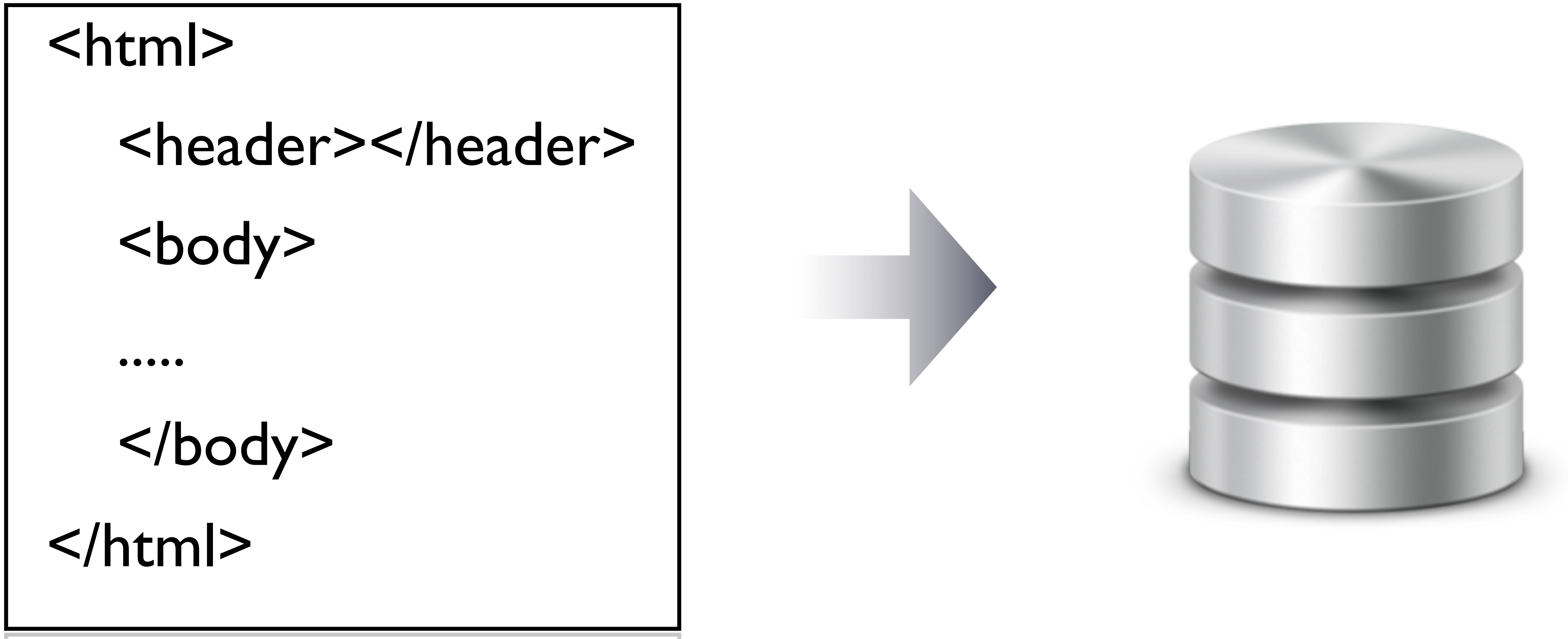
```
<?xml version="1.0" e
<xsl:stylesheet xmlns
<!-- created 2005-12-12-->
  <xsl:include href="xslt_
  <xsl:output method="xml"
  <xsl:template match="/">
  <root>
   Heuristic:<xsl:value-of
   <p>The leading manufact
  </root>
  </xsl:template>
 </xsl:stylesheet>
```

jQuery
mobile framework.

Plone®

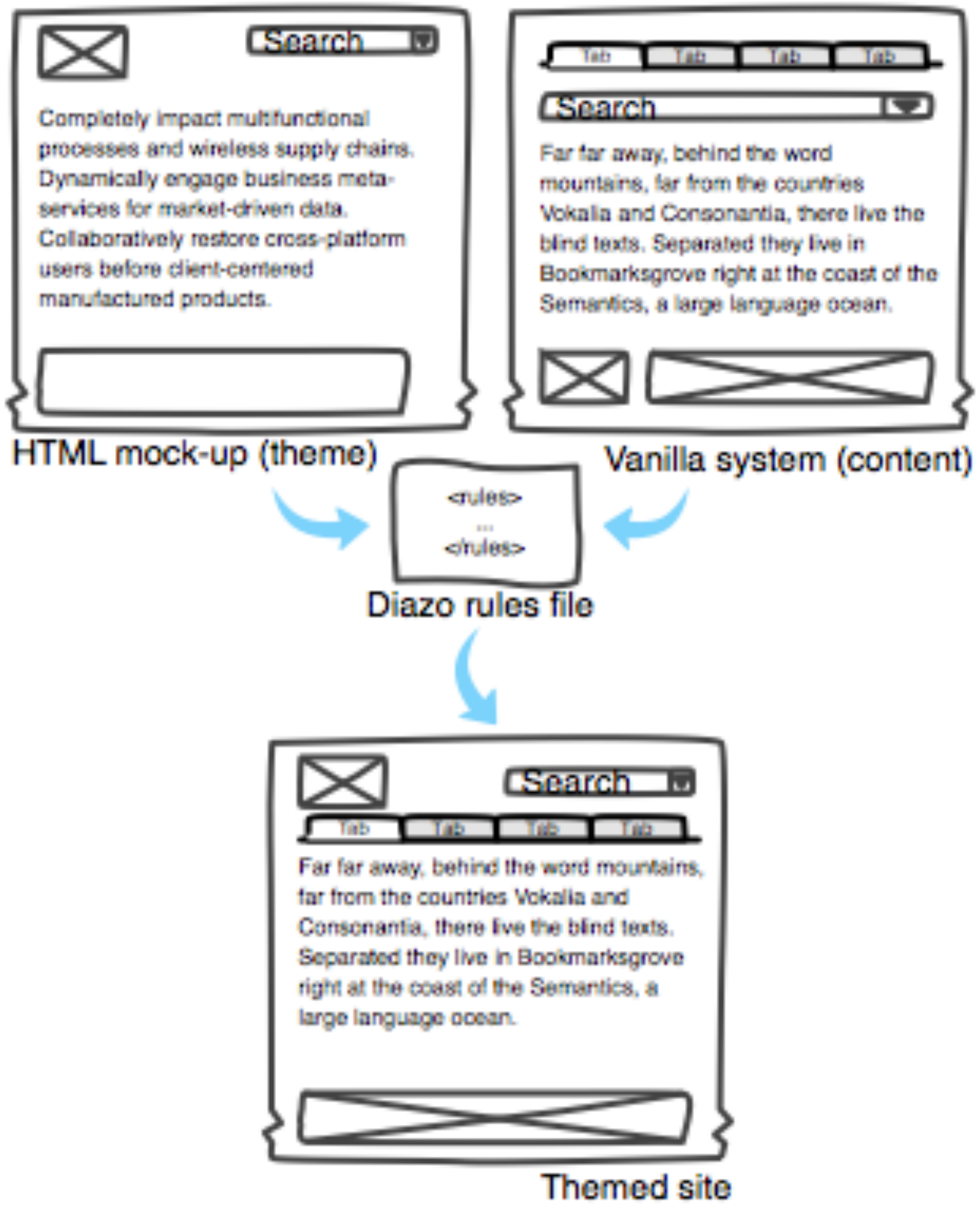
Data Scraping
vs
Web Scraping

Data Scraping



```
<html>
    <header></header>
    <body>
    .....
    </body>
</html>
```

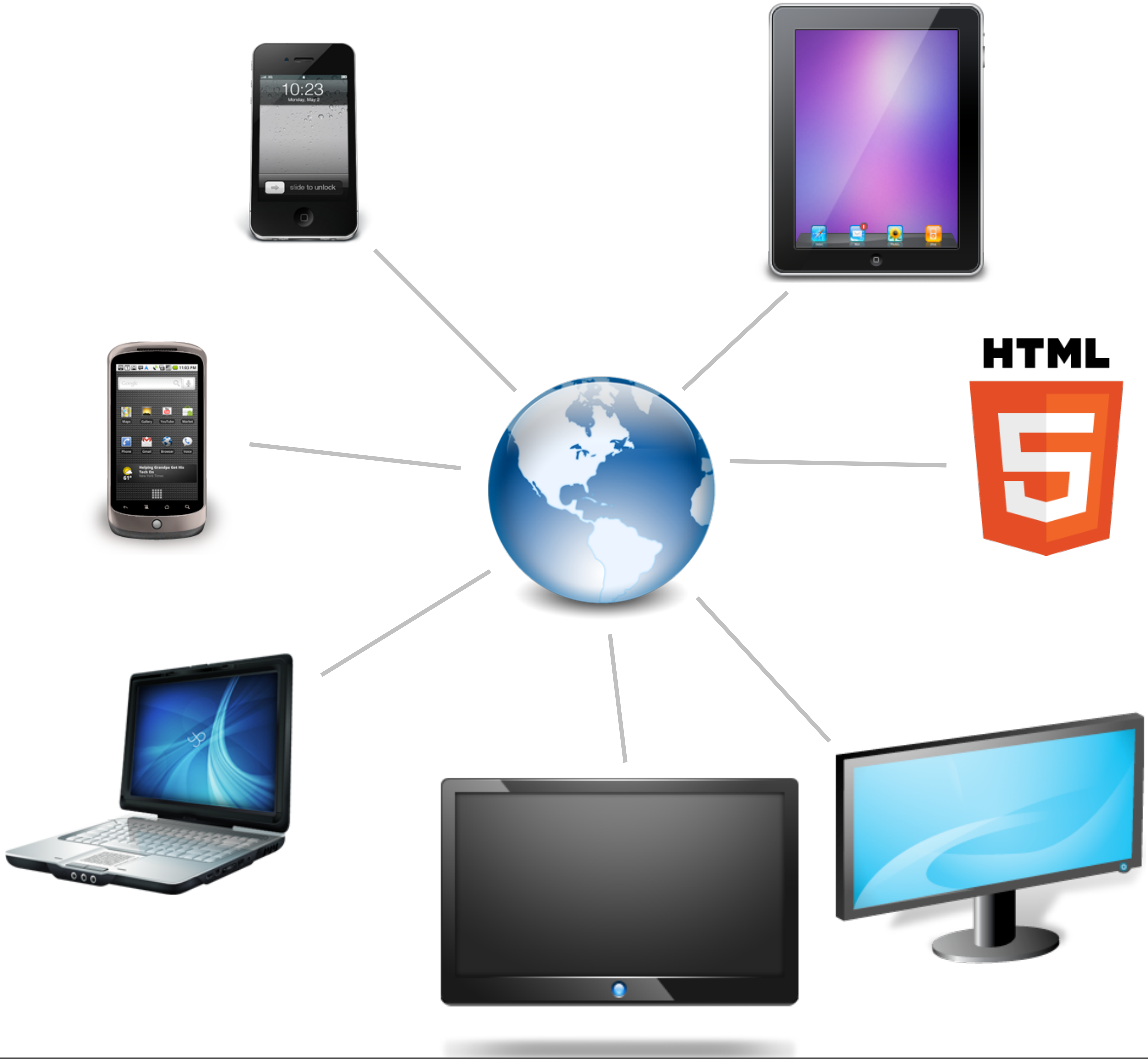
Web Scraping

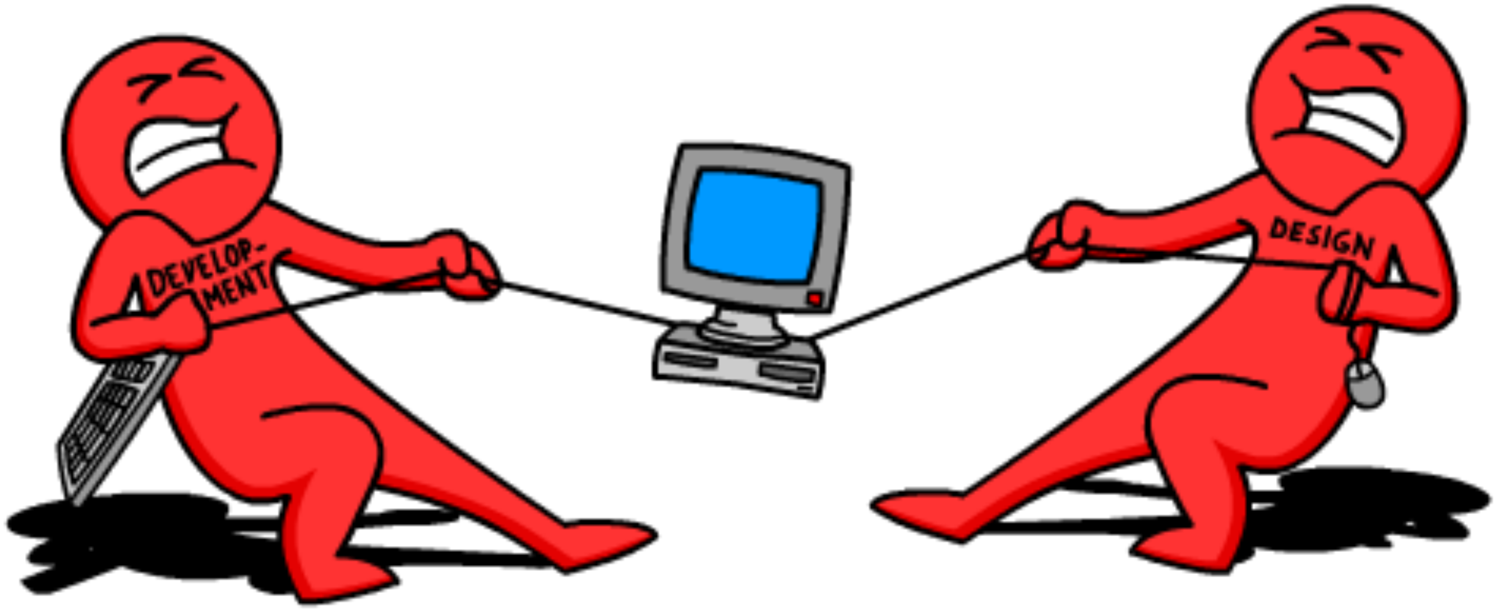


Completely impact multifunctional processes and wireless supply chains. Dynamically engage business meta-services for market-driven data. Collaboratively restore cross-platform users before client-centered manufactured products.

Far far away, behind the word mountains, far from the countries Vokalia and Consonantia, there live the blind texts. Separated they live in Bookmarksgrove right at the coast of the Semantics, a large language ocean.

Far far away, behind the word mountains, far from the countries Vokalia and Consonantia, there live the blind texts. Separated they live in Bookmarksgrove right at the coast of the Semantics, a large language ocean.


HTML
5



- "Why does it always take you so long to draw those damn boxes?"
- "For the last time, they're not boxes. They're called **divs**!"

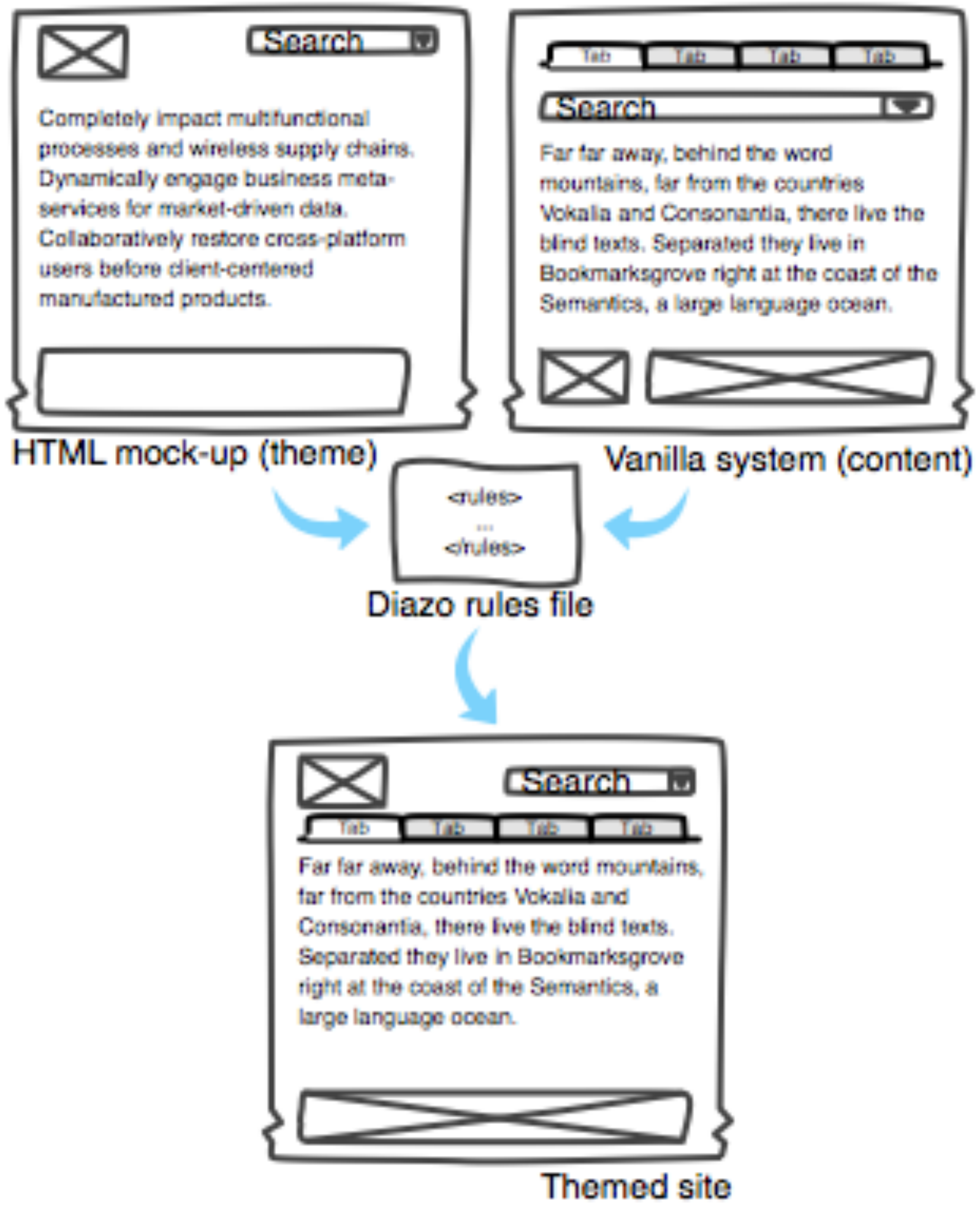
© 2010 Slekx.com

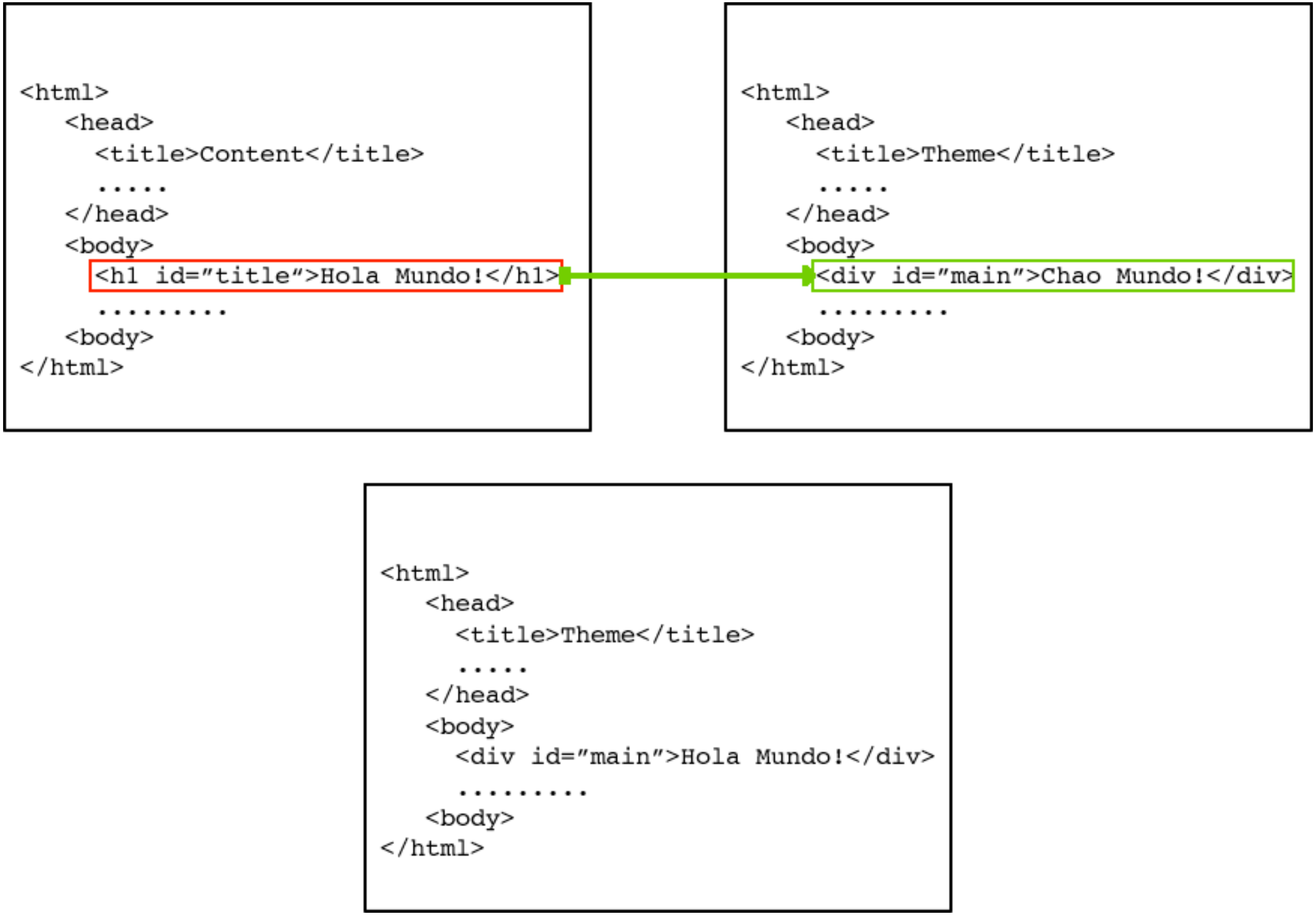
Deliverance

XDV

Diazo

Diazo


Search
Completely impact multifunctional processes and wireless supply chains. Dynamically engage business meta-services for market-driven data. Collaboratively restore cross-platform users before client-centered manufactured products.
HTML mock-up (theme)
Tab
Tab
Tab
Tab
Search
Far far away, behind the word mountains, far from the countries Vokalia and Consonantia, there live the blind texts. Separated they live in Bookmarksgrove right at the coast of the Semantics, a large language ocean.
Vanilla system (content)
<rules>
...
</rules>
Diazo rules file
Search
Tab
Tab
Tab
Tab
Far far away, behind the word mountains, far from the countries Vokalia and Consonantia, there live the blind texts. Separated they live in Bookmarksgrove right at the coast of the Semantics, a large language ocean.
Themed site



```
<html>
   <head>
     <title>Content</title>
     .....
   </head>
   <body>
     <h1 id=”title“>Hola Mundo!</h1>
     .........
   <body>
</html>
```

```
<html>
   <head>
     <title>Theme</title>
     .....
   </head>
   <body>
     <div id=”main”>Chao Mundo!</div>
     .........
   <body>
</html>
```

```
<html>
   <head>
     <title>Theme</title>
     .....
   </head>
   <body>
     <div id=”main”>Hola Mundo!</div>
     .........
   <body>
</html>
```

```
<replace css:content=”h1” css:theme=”#main” />
```

```
<drop css:content=”h1” />

<drop css:theme=”breadcrumbs” />
```

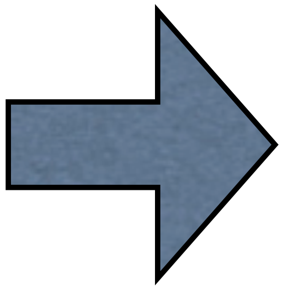
```
<replace css:theme=”#header” content=”#header-element” if-content=”” />
```

```
<drop css:theme="#info-box" if-path="/news"/>
```

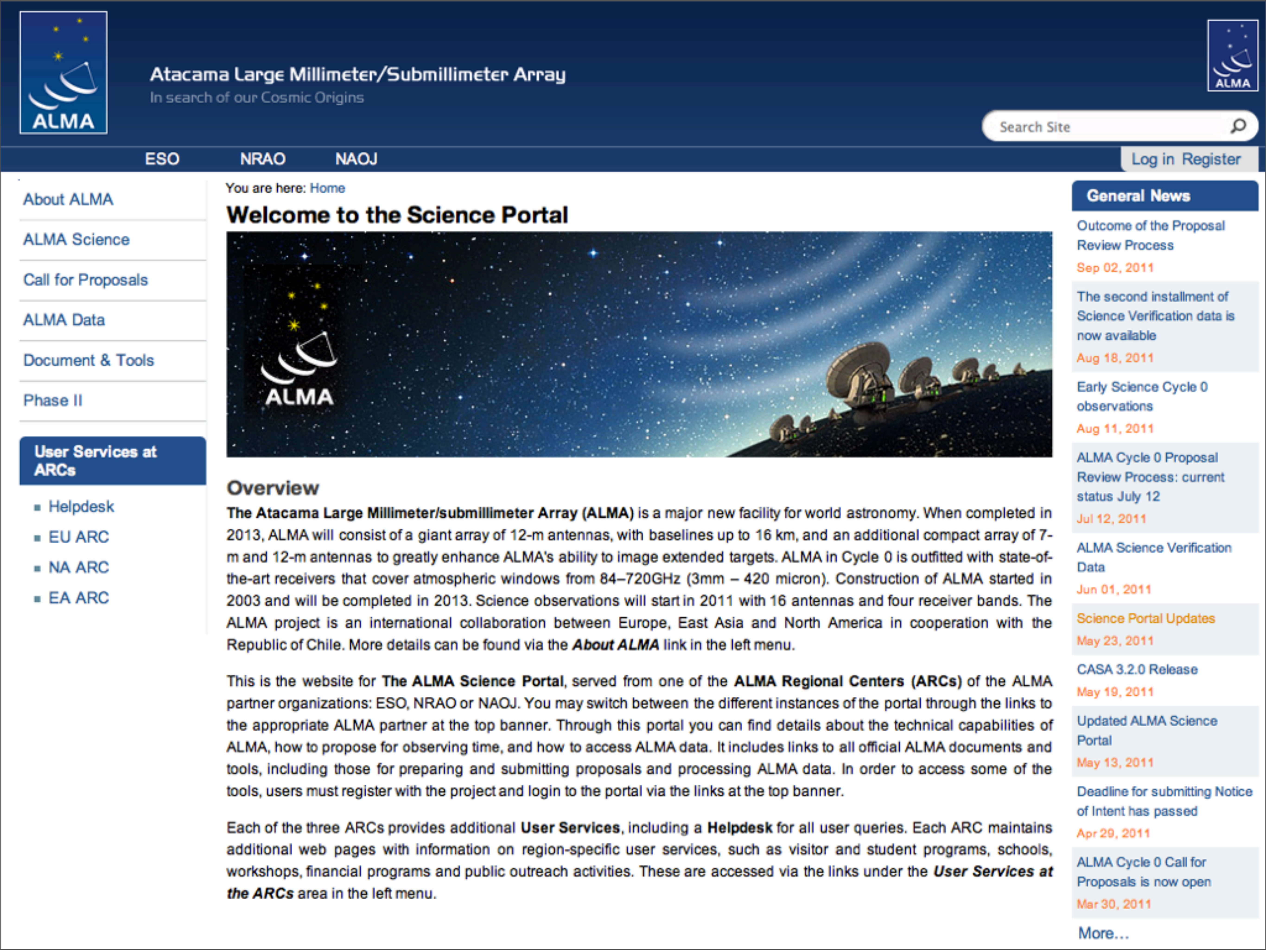
```
<theme/>
<notheme/>
<replace/>
<before/>
<after/>
<drop/>
<strip/>
<merge/>
<copy/>
```

```
<replace css:theme="#details">
   <dl id="details">
     <xsl:for-each css:select="table#details > tr">
       <dt><xsl:copy-of select="td[1]/text()" /></dt>
       <dd><xsl:copy-of select="td[2]/node()"/></dd>
     </xsl:for-each>
   </dl>
</replace>/></dt>
```

```
<table id="details">
    <tr>
        <td>One</td>
        <td>1</td>
    </tr>
    <tr>
        <td>Two</td>
        <td>2</td>
    </tr>
</table>
```



```
<dl id="details">
    <dt>One</dt>
    <dd>1</dd>
    <dt>Two</dt>
    <dd>2</dd>
</dl>
```



You are here: Home

Welcome to the Science Portal

Overview

The Atacama Large Millimeter/submillimeter Array (ALMA) is a major new facility for world astronomy. When completed in 2013, ALMA will consist of a giant array of 12-m antennas, with baselines up to 16 km, and an additional compact array of 7-m and 12-m antennas to greatly enhance ALMA's ability to image extended targets. ALMA in Cycle 0 is outfitted with state-of-the-art receivers that cover atmospheric windows from 84–720GHz (3mm – 420 micron). Construction of ALMA started in 2003 and will be completed in 2013. Science observations will start in 2011 with 16 antennas and four receiver bands. The ALMA project is an international collaboration between Europe, East Asia and North America in cooperation with the Republic of Chile. More details can be found via the ***About ALMA*** link in the left menu.

This is the website for **The ALMA Science Portal**, served from one of the **ALMA Regional Centers (ARCs)** of the ALMA partner organizations: ESO, NRAO or NAOJ. You may switch between the different instances of the portal through the links to the appropriate ALMA partner at the top banner. Through this portal you can find details about the technical capabilities of ALMA, how to propose for observing time, and how to access ALMA data. It includes links to all official ALMA documents and tools, including those for preparing and submitting proposals and processing ALMA data. In order to access some of the tools, users must register with the project and login to the portal via the links at the top banner.

Each of the three ARCs provides additional **User Services**, including a **Helpdesk** for all user queries. Each ARC maintains additional web pages with information on region-specific user services, such as visitor and student programs, schools, workshops, financial programs and public outreach activities. These are accessed via the links under the ***User Services at the ARCs*** area in the left menu.

General News

Outcome of the Proposal Review Process
Sep 02, 2011

The second installment of Science Verification data is now available
Aug 18, 2011

Early Science Cycle 0 observations
Aug 11, 2011

ALMA Cycle 0 Proposal Review Process: current status July 12
Jul 12, 2011

ALMA Science Verification Data
Jun 01, 2011

Science Portal Updates
May 23, 2011

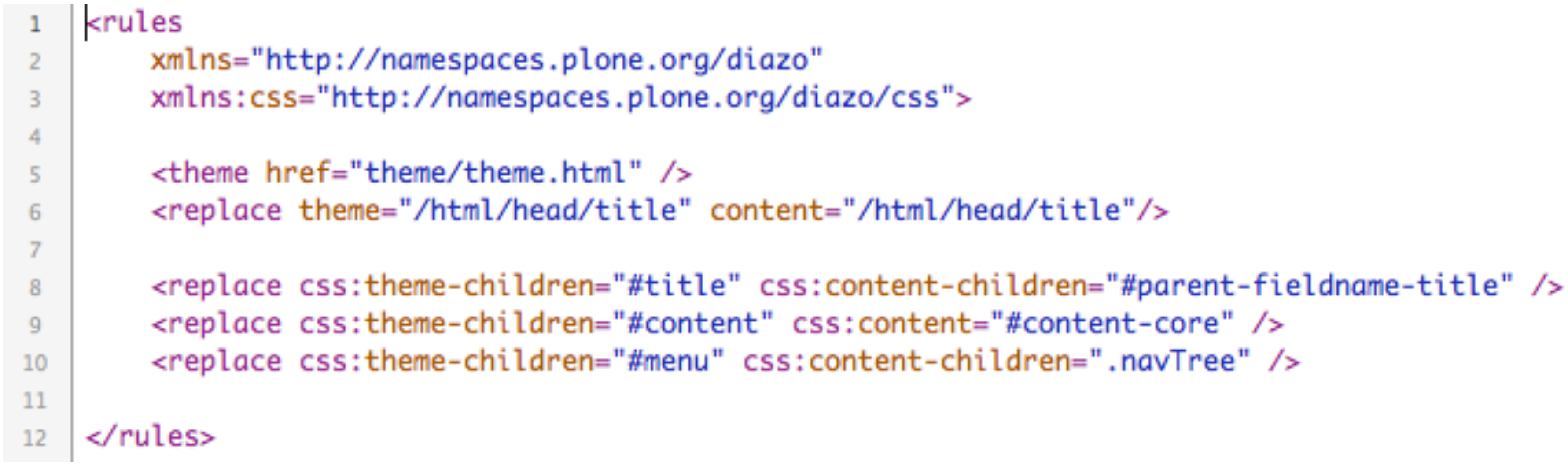
CASA 3.2.0 Release
May 19, 2011

Updated ALMA Science Portal
May 13, 2011

Deadline for submitting Notice of Intent has passed
Apr 29, 2011

ALMA Cycle 0 Call for Proposals is now open
Mar 30, 2011

More...

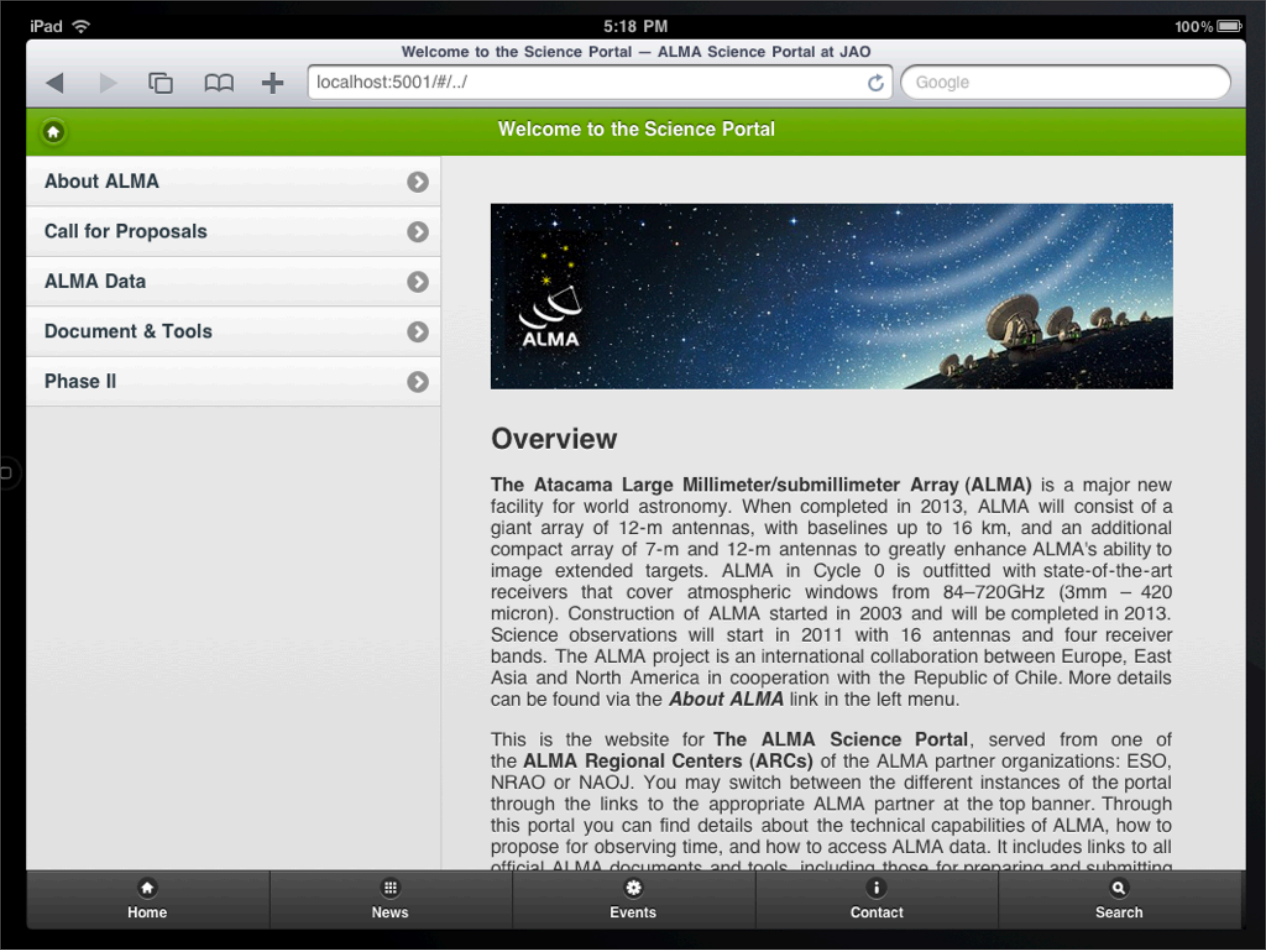


```xml
<rules
    xmlns="http://namespaces.plone.org/diazo"
    xmlns:css="http://namespaces.plone.org/diazo/css">

    <theme href="theme/theme.html" />
    <replace theme="/html/head/title" content="/html/head/title"/>

    <replace css:theme-children="#title" css:content-children="#parent-fieldname-title" />
    <replace css:theme-children="#content" css:content="#content-core" />
    <replace css:theme-children="#menu" css:content-children=".navTree" />

</rules>
```



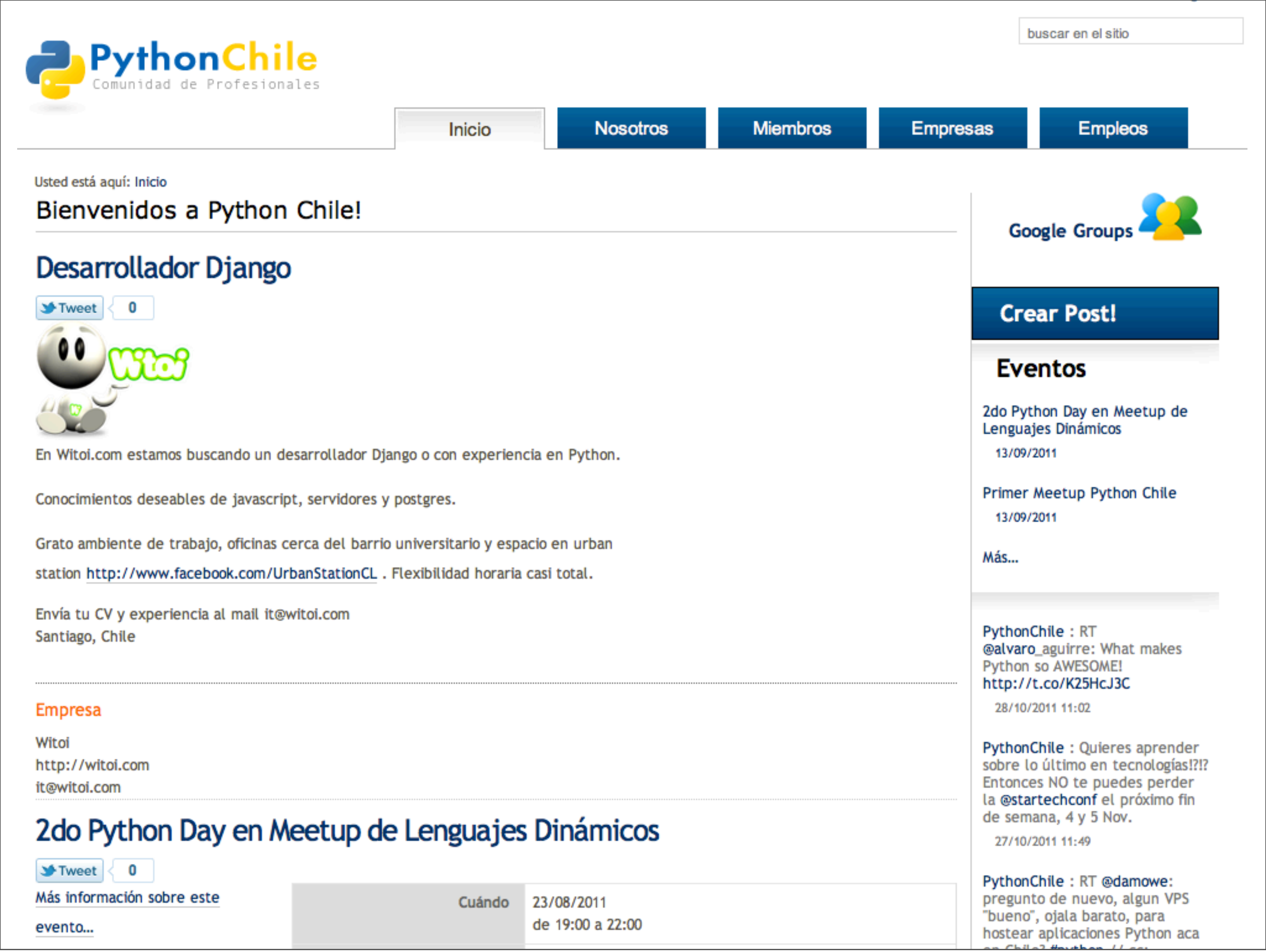
Overview

The Atacama Large Millimeter/submillimeter Array (ALMA) is a major new facility for world astronomy. When completed in 2013, ALMA will consist of a giant array of 12-m antennas, with baselines up to 16 km, and an additional compact array of 7-m and 12-m antennas to greatly enhance ALMA's ability to image extended targets. ALMA in Cycle 0 is outfitted with state-of-the-art receivers that cover atmospheric windows from 84–720GHz (3mm – 420 micron). Construction of ALMA started in 2003 and will be completed in 2013. Science observations will start in 2011 with 16 antennas and four receiver bands. The ALMA project is an international collaboration between Europe, East Asia and North America in cooperation with the Republic of Chile. More details can be found via the ***About ALMA*** link in the left menu.

This is the website for **The ALMA Science Portal**, served from one of the **ALMA Regional Centers (ARCs)** of the ALMA partner organizations: ESO, NRAO or NAOJ. You may switch between the different instances of the portal through the links to the appropriate ALMA partner at the top banner. Through this portal you can find details about the technical capabilities of ALMA, how to propose for observing time, and how to access ALMA data. It includes links to all official ALMA documents and tools, including those for preparing and submitting

Tools

External Content



Usted está aquí: Inicio

Bienvenidos a Python Chile!

Desarrollador Django

En Witoi.com estamos buscando un desarrollador Django o con experiencia en Python.

Conocimientos deseables de javascript, servidores y postgres.

Grato ambiente de trabajo, oficinas cerca del barrio universitario y espacio en urban station http://www.facebook.com/UrbanStationCL . Flexibilidad horaria casi total.

Envía tu CV y experiencia al mail it@witoi.com
Santiago, Chile

Empresa

Witoi
http://witoi.com
it@witoi.com

2do Python Day en Meetup de Lenguajes Dinámicos

Más información sobre este evento...

Cuándo	23/08/2011 de 19:00 a 22:00

Eventos

2do Python Day en Meetup de Lenguajes Dinámicos
13/09/2011

Primer Meetup Python Chile
13/09/2011

Más...

PythonChile : RT @alvaro_aguirre: What makes Python so AWESOME! http://t.co/K25HcJ3C
28/10/2011 11:02

PythonChile : Quieres aprender sobre lo último en tecnologías!?!? Entonces NO te puedes perder la @startechconf el próximo fin de semana, 4 y 5 Nov.
27/10/2011 11:49

PythonChile : RT @damowe: pregunto de nuevo, algun VPS "bueno", ojala barato, para hostear aplicaciones Python aca

- development of web & mobile interfaces
- legacy apps integrations
- prototypes
- low coupling

```python
from diazo.compiler import compile_theme
from lxml import etree
from diazo.compiler import compile_theme

absolute_prefix = "/static"

rules = "rules.xml"
theme = "theme.html"

compiled_theme = compile_theme(rules, theme,
                               absolute_prefix=absolute_prefix)

transform = etree.XSLT(compiled_theme)
content = etree.parse(some_content)
transformed = transform(content)

output = etree.tostring(transformed)
```

github/aaguirre

diazo.org

plone.org

gracias!