



ACAP

Implementation Webinar

8th May 2008
17:00 BST

*For dial-in details see separate email
or*

*International access: +44 207 154 2640
or +44 207 769 6472
Pass code: 9563212*



ACAP Implementation Webinar

8th May 2008 – 17:00 BST

Welcome...

Aims of the Webinar:

- to give a basic introduction to ACAP
 - what is ACAP?
 - why you should consider implementing ACAP
 - how to implement ACAP in a simple way
 - how ACAP affects web crawlers now
- to introduce some of ACAP's special features
 - features that ACAP provides now
 - how to implement them
 - other features to be introduced in future



ACAP Implementation Webinar

What is ACAP?

ACAP is...

- a “toolkit” for communicating access and usage policies for your online content
 - a “policy” is a set of statements of how you wish your content to be used by others, such as by search engines
 - do you want your content to be crawled and copied by search engines and others?
 - if so, can they do whatever they like with your content, or do you wish to constrain how they use it?
 - for communication to whom?
 - to search engines and others using automated crawlers to take copies of your online content for use in their commercial products and services



ACAP Implementation Webinar

What is ACAP?

What is in the ACAP “toolkit”?

- A *dictionary* of terminology to be used in expressing access and usage policies
- A *technical specification* of how to include ACAP policy expressions in **robots.txt** and in **Robots <META> Tags**
 - extension of the Robots Exclusion Protocol (REP)
- An ACAP *Implementation Guide*
- An *online tool* for converting your existing robots.txt file to implement ACAP “simply”
- Guidance on crawler authentication



ACAP Implementation Webinar

What is ACAP?

ACAP is NOT...

- a formal standard
 - although it may become one
- a technical enforcement mechanism
 - it enables you to **express** your policy, but not **enforce** it
 - a crawler can choose either to act in accordance with your expressed policy, or ignore it



ACAP Implementation Webinar

Why implement ACAP?

Some reasons why your business might wish to implement ACAP...

- You are losing traffic to your website, because your readers can find copies of your content elsewhere, such as in search results
- As a result of poor “footfall” on your website, you are losing the advertising revenue that you need to maintain your web presence
- You see marketing opportunities in making paid-for content visible to crawlers, but need to protect your revenue



ACAP Implementation Webinar

Why implement ACAP?

What implementing ACAP will do for you...

- It enables you to express your access and usage policies in a *machine-readable format*
- It sends a signal to crawler operators that you care how they use your content
- Without ACAP, you are strictly limited in the policies that can be expressed...
 - REP only contains simple forms of expression that suit the search engines
- ...but with ACAP, you can express your policy more precisely
 - even if it doesn't suit the search engines



ACAP Implementation Webinar

Why implement ACAP?

Why have we used REP?

- The Robots Exclusion Protocol is the most widely-used method for communicating access and usage policies to crawlers
 - millions of websites have a robots.txt file
- The search engines recommend REP for communicating with their crawlers
- Otherwise it would not have been our choice
 - other formats (e.g. sitemaps, other XML formats) would have suited our purpose better



ACAP Implementation Webinar

Getting started...

First, implement ACAP in a simple way...

- Convert your robots.txt file to use ACAP syntax, using the online tool
 - All the existing content of your robots.txt file is still there, unaltered
 - the-acap.org/convert-robots-txt-to-acap.php
- Put the ACAP logo on your website home page
 - the-acap.org/add-acap-enabled.php



ACAP Implementation Webinar

Getting started...

If you don't have a robots.txt file for your website, can you implement ACAP?

- Yes, but having a robots.txt file makes it easy to implement ACAP in a simple way
- You can implement ACAP purely using embedded Robots <META> Tags
 - for some CMS-managed websites, this may be appropriate
 - but you will need to discuss this with your CMS vendor
 - *Atex* and *escenic* are both supporting ACAP



ACAP Implementation Webinar

Getting started...

What your robots.txt file may look like now...

```
# My robots.txt file
# comments are ignored by crawlers...

# Deny access to crawler robot badbot
User-agent: badbot
Disallow: /

# Ask others not to crawl /temp/
User-agent: *
Disallow: /temp/
```



ACAP Implementation Webinar

Getting started...

What the online tool will add...

```
# My robots.txt file
# comments are ignored by crawlers...
# Deny access to crawler 'badbot'
ACAP-crawler: badbot
# User-agent: badbot
ACAP-disallow-crawl: /
# Disallow: /
# Ask others not to crawl /temp/
ACAP-crawler: *
# User-agent: *
ACAP-disallow-crawl: /temp/
# Disallow: /temp/
```



ACAP Implementation Webinar

Getting started...

Why duplicate what is already in robots.txt?

- There is no single, standard interpretation of conventional REP
 - Does 'Disallow' mean “don't crawl” or “don't index”?
 - In ACAP 'disallow-crawl' means “don't crawl”!
- It is important to avoid any possibilities of ambiguity in the interpretation of robots.txt
 - Maintain conventional REP expressions for crawlers that continue not to implement ACAP
 - Maintain ACAP equivalent expressions, to be ready when crawlers start to interpret ACAP expressions



ACAP Implementation Webinar

Getting started...

Let's try it...

the-acap.org/convert-robots-txt-to-acap.php

the-acap.org/add-acap-enabled.php



ACAP Implementation Webinar

How ACAP affects crawlers now

What will happen when you implement ACAP?

- No major crawler operators have currently implemented ACAP
 - Their crawlers have not been programmed to recognise and interpret ACAP expressions
 - Crawlers are designed to ignore anything in robots.txt that they are not programmed to recognise as conventional REP expressions
 - Why? Because with millions of mostly hand-crafted robots.txt files out there, many thousands contain errors
- There is no evidence that any crawlers have yet changed their behaviour because of ACAP



ACAP Implementation Webinar

How ACAP affects crawlers now

How many websites have implemented ACAP?

- The number is small but growing...
 - currently approximately 150 websites
 - across more than 20 countries
 - mostly the websites of publishers with high-value content
- *No-one is taking any risks...*
- *...because there are no risks*



ACAP Implementation Webinar

Taking ACAP to the next level

In summary...

- ACAP enables you to express more than conventional REP
- A simple ACAP implementation only expresses what you have already expressed in your existing robots.txt file
- You can also add ACAP policy expressions directly to HTML pages
- To implement ACAP more fully, we recommend that you read the ACAP Implementation Guide
http://the-acap.org/download.php?ACAP-TF-CrawlerCommunication-ImplementationGuide_V1%20_ISSUE_1.pdf
- Full technical specifications are also available from the ACAP website



ACAP Implementation Webinar

Taking ACAP to the next level

What more can ACAP do?

- ACAP enables you to express policies about how your content is accessed (crawled) and used
- ACAP uses its own vocabulary, as defined in the ACAP Dictionary, to encourage consistent interpretation by implementers
- We have already met one ACAP term: **crawl**
- There are five other major terms for the various types of usage that apply to search engines:
 - **follow, index, preserve, present, other**



ACAP Implementation Webinar

Taking ACAP to the next level

What do these usage types mean?

- ***crawl*** = request a copy of this page / resource
- ***follow*** = look for links in the page and follow them to find other pages / resources
- ***index*** = index a resource so that it can be found in search results
- ***preserve*** = store for an indefinite period a cached copy of the page / resource
- ***present*** = present the resource to an end-user, whether in full or as a snippet, thumbnail, etc.
- ***other*** = use the resource in any way other than specifically permitted
 - used in prohibitions only



ACAP Implementation Webinar

The basic ACAP usage types

- Each type of usage can be permitted or prohibited using an ACAP expression, e.g.
 - ACAP-allow-crawl...
 - ACAP-disallow-crawl...
 - ACAP-allow-index...
 - ACAP-disallow-index...
 - ACAP-allow-follow...
 - ACAP-disallow-follow...
 - ...



ACAP Implementation Webinar

The basic ACAP usage types

But doesn't REP support 'index' and 'follow'?

- Yes, but only in Robots <META> Tags
- ACAP adds the ability to express 'index' and 'follow' permissions in robots.txt
- ACAP supports policy expression both in robots.txt and in Robots <META> Tags, e.g.

- in robots.txt

```
ACAP-allow-index: /public/*.htm
```

- in an HTML <META> tag:

```
<META name="robots" content="ACAP allow-index">
```



ACAP Implementation Webinar

More usage types

- In addition to the basic usage types, there are additional usage types for presenting resources in different ways:
 - present-snippet
 - present-thumbnail
 - present-currentcopy
 - present-oldcopy
 - present-original
 - present-link



ACAP Implementation Webinar

Expressing qualified permissions

- ACAP allows you to qualify a permission in various ways, according to the usage
 - Permission to index, preserve or present can be qualified by a time limit, e.g.

```
ACAP-allow-index /news/ time-limit=until-recrawled
```

```
ACAP-allow-preserve /news/ time-limit=until-2008-06-30
```

```
ACAP-allow-present-snippet /news/ time-limit=30-days
```

- Permission to present a snippet can be qualified by a length limit, e.g.

```
ACAP-allow-present-snippet /news/ max-length=250-chars
```



ACAP Implementation Webinar

Expressing qualified permissions

- There are more ways of qualifying permissions...
 - Permission to present a preserved copy may be qualified by prohibiting certain kinds of modification, e.g.

`ACAP-allow-present-currentcopy /news/ prohibited-modification=translation`

- Other qualifiers are described in the ACAP Version 1.0 technical specifications, available from the ACAP website



ACAP Implementation Webinar

Usage purposes: why crawlers crawl

- ACAP recognises that some crawlers crawl for a variety of purposes
 - to feed a general-purpose search service
 - e.g. Google Web Search
 - to feed a special-purpose search service
 - e.g. Yahoo! Images Search
 - to feed a search service based on a specific country
 - e.g. fr.msn.com
 - to re-syndicate content to commercial partners
- Using ACAP you can communicate policies to a crawler that are specific to a particular usage purpose, e.g.
 - `ACAP-usage-purpose: google.fr`
 - `ACAP-disallow-index: / # don't index in google.fr`



ACAP Implementation Webinar

Embedding policies in HTML pages

- ACAP policies can be embedded in HTML pages using Robots <META> Tags

```
<META name="robots" content="...">
```

- ACAP policies contained in <META> Tags apply to the whole page, e.g.

```
<META name="robots"  
content="ACAP google.fr allow-present-snippet  
time-limit=30-days max-length=100-words">
```



ACAP Implementation Webinar

Putting it all together...

- A complete set of policies in robots.txt will contain a set of ACAP records
 - each record addressed to one or more crawlers
 - each record possibly sub-divided by usage purpose
 - each record containing as many permissions and prohibitions
 - for different sets of pages and resources
 - for whichever usages are relevant



ACAP Implementation Webinar

Putting it all together...

- The ACAP Implementation Guide provides a step-by-step guide to deciding how to express your policies
 - in robots.txt
 - in Robots <META> Tags
 - with lots of examples

http://the-acap.org/download.php?ACAP-TF-CrawlerCommunication-ImplementationGuide_V1%20_ISSUE_1.pdf



ACAP Implementation Webinar

Plans for additional features

- ACAP isn't only about search engines
- ACAP is about all forms of machine-to-machine communication of access and usage policies
- ...but search engines are pretty important



ACAP Implementation Webinar

Plans for additional features

- Future revisions of ACAP will include:
 - some specific features for communicating with web archive crawlers
 - new ways of communicating with crawlers, e.g.
 - method for embedded policies in PDF, images and other media resources
 - support for expressing permissions for fragments of web pages
 - a crawler description language
 - to enable you to find out everything you need to know about a crawler
 - such as: whether or not it can interpret ACAP



ACAP Implementation Webinar

Finding out more...

- We hope this has been a useful introduction
- To find out more, please visit our website

<http://the-acap.org>

- and contact us if you need further help

info@the-acap.org



ACAP Implementation Webinar

That's all, folks!

and we wish you every success with
your ACAP implementation!