

ACAP Pilot: Summary of use cases being tested

Francis Cave, ACAP Technical Project Manager

Public document, 16 September 2007

Introduction

At the start of the ACAP pilot project, each of the publisher participants were invited to prepare use cases for their requirements for the ACAP Technical Framework.

These use cases have been vital in determining the initial content of the ACAP Technical Framework. They have ensured that the Technical Framework contains components that meet genuine publisher requirements for the communication of permissions for access and use of online content.

Having reviewed and discussed the content of each of these use cases, and prepared drafts of the necessary components of the Technical Framework, the project has now entered the testing phase when these components are being implemented and tested by several of the participants to ensure that they not only meet the requirements of the use cases but are also feasible to implement.

An analysis of the use cases carried out earlier in the project determined that the principal components of the Technical Framework that are needed to meet the requirements of these initial uses cases are:

- Extension of the Robots Exclusion Protocol to enable ACAP permissions and prohibitions relating to an entire site to be communicated to crawlers, using 'robots.txt'.
- Extension of the Robots Exclusion Protocol to enable ACAP permissions and prohibitions relating to an individual content item to be embedded in HTML pages, using Robots META Tags.
- A mechanism for embedding permissions in HTML page fragments, using the HTML 'class' attribute.
- A mechanism for embedding permissions in non-text content items, such as PDF and image files.
- ACAP recommendations of web server procedures for identifying and authenticating visiting crawlers, so that privileged crawlers may be given access to premium content protected by firewalls.
- ACAP recommendations of how to request that search engines and other aggregators 'take down' content items, by removing entries from their indexes and all copies from internal stores, in order to comply with actual or anticipated legal action (e.g. in the case of libellous or other factually incorrect items).
- A mechanism for including ACAP permissions in XML-based communication formats, such as are used for information feeds and content syndication.

The first three of these requirements are intended to be met by components of the ACAP Technical Framework that have been published so far, and the remaining requirements will be met by further components to be published before the end of this phase of the project.

The following summarises the use cases that are the primary focus of implementation and testing of the ACAP Technical Framework.

1 Macmillan/Holtzbrinck: access to firewall-protected online book content with a worldwide readership

Verlagsgruppe Georg von Holtzbrinck is a global publishing group within interests across news, magazine, journal and book publisher, including Macmillan.

Macmillan, apart from being a leading publisher of printed books and journals, is also a successful online publisher and, through a subsidiary company MPS Technologies, provider of online services to other publishers.

1.1 Use cases to be tested

The BookStore platform provides publishers with a secure means of putting book content online. Macmillan need a means of giving privileged search engine crawlers access to book content that is not openly accessible, so that paid-for book content can be indexed by the search engines, thereby greatly improving the visibility of online books to the consumer without compromising the security of the content.

Macmillan will test that the ACAP Technical Framework meets the following use cases (using the proposed extensions to the Robots Exclusion Protocol):

- Communication of which crawlers and services associated with a specific crawler may index and provide links to their content.
- Communication of what they wish a search result to show, including indicators on end-user access and use of book content (for example, that end-users will have to pay to get access to the entire book).

Macmillan will also test the following use cases, when the necessary components of the ACAP Technical Framework are ready:

- Authentication of crawlers visiting their site, so that privileged crawlers may be given access for indexing purposes to paid-for book content stored behind a firewall.
- Communication of permission embedded in book content in a variety of formats.

1.2 Technical approach

Macmillan is implementing a pilot server on which to host test content and to test crawler authentication techniques.

2 Media 24: access to current and archived online news published in South Africa

Media 24 publishes a wide range of print and online content, including magazines and books, in South Africa and elsewhere in Africa and Asia. In South Africa they

publish six national daily newspapers, five weeklies and 48 regional newspapers. The magazine division publishes over 60 titles. The company's activities also include a variety of educational and book publishing. The online business of Media 24 attracts nearly a million unique visitors per month.

2.1 Use cases to be tested

Media 24 will test use cases in the news publishing section of their business. They will test the proposed extensions to the Robots Exclusion Protocol, and especially Part 2 for embedding permissions in individual HTML pages and for associating permissions with page fragments.

Media 24 will test that the following use cases are met:

- Communication of which crawlers and services associated with a specific crawler may crawl a website and present their content.
- Communication of which links crawlers and services may follow.
- Communication of which crawlers and services associated with a specific crawler may index with specific reference to restrictions on video and sound.
- Communication of how long crawlers and services may store copied content.
- Communication of how crawlers and services should render content with specific reference to cached versions.
- Communication of what other permissions and restrictions apply to crawlers and services.

2.2 Technical approach

Media 24 are implementing a pilot server on which to host test content. Permissions and restrictions will be hard coded for purposes of this phase of the pilot.

3 De Persgroep: access to current and archived online news published in Belgium

De Persgroep is a leading publisher of printed and online news in Flanders. Their news sites include **hln.be** with a monthly reach of 3 million visitors, **demorgen.be**, with a monthly reach of approximately 500,000, and **7s7.be**, a site for French-speaking Belgians with a monthly reach of approximately 750,000.

Access to much of De Persgroep's online content is free, but in the near future some content will be paid-for and admittance will only be allowed to paying subscribers.

3.1 Use cases to be tested

De Persgroep will test the proposed extensions to the Robots Exclusion Protocol, and especially Part 1 Extensions to robots.txt, to ensure that the follow use cases are met:

- Communication of which crawlers and services associated with a specific crawler may index, store (archive) and present their content.
- Communication of (confirmations of) requests to take down pages, i.e. remove them from search engine indexes.
- Communication of constraints on what a search result for their content should contain, including specified text in snippets and snippet length.

De Persgroep will also test the following use cases, when the necessary components of the ACAP Technical Framework are ready:

- Authentication of crawlers visiting their site, so that privileged crawlers may be given access for indexing purposes to paid-for archive content stored behind a firewall.

3.2 Technical approach

De Persgroep are implementing a pilot server on which to host test content. The pilot server will test mechanisms for authenticating crawlers and giving them access to protected content, as well as testing the communication of permissions using robots.txt and embedded META Tags.

4 Reed Elsevier: access to online scholarly journal content with a worldwide readership

Reed Elsevier is a world-leading publisher of information for professional users. The Reed Elsevier group of companies includes Elsevier, a publisher of online science and health information, and Reed Business Publishing, which has significant online activity in business information publishing.

4.1 Use cases to be tested

Reed Elsevier will test use cases in both scientific and business information publishing. Reed Elsevier will test the proposed extensions to the Robots Exclusion Protocol, especially the ability to communicate permissions to index content based upon:

- the search services for which content will appear in search results
- whether the content has been harvested from a legitimate location.

Reed Elsevier will also test the ability to communicate what search results are allowed to contain.

When the necessary components of the ACAP Technical Framework are ready, Reed Elsevier will also test the ability to communicate ACAP permissions embedded within PDF files.

4.2 Technical approach

Reed Elsevier will supply content and permissions (both embedded within content and in 'robots.txt' format) for uploading to a shared ACAP server.