

# ACAP Technical Framework: Usage Definitions

## Draft for pilot testing, Version 2

For the purposes of the ACAP Pilots, six basic usage verbs are now proposed: crawl, follow, index, store, present, other. Each is specified in tabular form, with additional notes, in sections 1 to 6 below. Section 7 is a glossary of terms used. The first appearance of each term defined in the glossary is shown in blue.

This paper is explicitly limited to usages that may be required for Pilot use cases. Other usages will need to be considered at later stages of ACAP development. Qualifier values that are expected to be tested in Pilot use cases are **highlighted thus**. Some other qualifiers and qualifier values are also included, but these should not, of course, be regarded as exhaustive.

### 1. Crawl

Usage	Crawl
Definition	Retrieve a <b>content item</b> for processing in accordance with other permissions and prohibitions.
User	<b>Crawler</b>
Used Resource	Content item
Used Resource Qualifiers	None
Output Resource	<b>Content item retrieved copy</b>
Usage Qualifiers	None
Dependencies	None
Notes	None

### 2. Follow

Usage	Follow
Definition	Follow links contained in a content item, except where explicitly prohibited at a more detailed level.
User	Crawler
Used Resource	Content item retrieved copy
Used Resource Qualifiers	None
Usage Qualifiers	None
Dependencies	“Crawl” must be permitted.
Note	A prohibition on “follow” is not to be interpreted as prohibiting any other use of links, eg in page ranking.

## ACAP Technical Framework – Public draft

### 3. Index

Usage	Index
Definition	Derive and store <a href="#">content item index entries</a> in accordance with a service operator's indexing practices
User	<a href="#">Process</a>
Used Resource	Content item retrieved copy, <i>or</i> Content item retrieved copy specified extract, <i>or</i> Publisher-specified indexing terms
Used Resource Qualifier (1)	Content property
Used Resource Qualifier (2)	Location (to prevent indexing of unauthorised copies) - index if found at: <a href="#">URI of authorised location</a> Repeatable if there is more than one authorised location (eg a mirror site).
Used Resource Qualifier (3)	Specified indexable content (if Used Resource is either Content item retrieved copy specified extract, <i>or</i> Publisher-specified indexing terms): <a href="#">URI of indexable content</a>
Output Resource	Content item index entries
Usage Purpose	<a href="#">Service</a>
Usage Qualifier	Time period: <a href="#">until recrawled</a> <a href="#">permanently</a> (eg for national library web archive) <a href="#">until date</a> for a specified period from a specified event
Dependencies	"Crawl" must be permitted.
Note (1)	If "index" is allowed for a specified page, it is to be interpreted as applying to the whole of the page, including links that are visible in the page but that are subject to a "no-follow" prohibition.
Note (2)	Permission to "index" does not imply permission to extract indexed content for "data mining" applications, which are not covered within the pilot set of usages.

**4. Preserve**

<b>Usage</b>	<b>Preserve</b>
Definition	Store a copy of a content item, preserving as far as possible to arrangement and style of presentation of the original
User	<a href="#">Process</a>
Used Resource	Content item retrieved copy
Used Resource Qualifier (1)	Content property
Used Resource Qualifier (2)	Location (to prevent storage of unauthorised copies) - preserve if found at: <a href="#">URI of authorised location</a> Repeatable if there is more than one authorised location (eg a mirror site).
Output Resource	<a href="#">Content item preserved copy</a>
Usage Purpose	<a href="#">Service</a>
Usage Qualifier	Time period: <a href="#">until recrawled</a> <a href="#">permanently</a> (eg for national library web archive) <a href="#">until date</a> for a specified period from a specified event
Dependencies	“Crawl” must be permitted.
Note (1)	The word “preserve” has been chosen in preference to “cache” or “archive”, both of which are currently used by some search engines in ways that are not necessarily consistent, and in preference to “store”, which is seen as too vague to be applicable.
Note (2)	Storage of index entries does not constitute the preservation of a copy, even if index entries could be used to re-create a version of the original content item.
Note (3)	If permission to preserve is denied, the content item retrieved copy is required to be discarded after indexing. For current search engines, this may mean that a “no-preserve” item will not be included in applicable service(s), since many search engines preserve the content item even when its use for display (“present”) is prohibited. The option to specify “no-preserve” is retained, however, since for some publishers and some types of material it may become meaningful in future.

5. Present

Usage	Present
Definition	Present as part of a service delivered to a browser
User	Process
Used Resource	Content item (ie the original content item on the publisher’s website) <i>and/or</i> Content item preserved copy <i>and/or</i> Content item preserved copy extract <i>and/or</i> Content item snippet <i>and/or</i> Content item thumbnail <i>and/or</i> Publisher content item summary <i>and/or</i> Content item link If a set of permissions includes more than one Used Resource for a Present usage, this is to be taken to indicate than the permitted types may be used singly or in combination.
Used Resource Qualifier (1)	Content property
Used Resource Qualifier (2)	Location (to prevent use of unauthorised copies) - present if found at: URI of authorised location Repeatable if there is more than one authorised location (eg a mirror site).
Used Resource Qualifier (3)	Maximum length (of snippet or extract): number of characters number of words
Used Resource Qualifier (4)	Preserved copy version: current (i.e. copy of most recent available version) old (i.e. copy of any other version)
Usage Purpose	Service
Usage Qualifier (1)	Modification applied to content item: none format (i.e. data format) style (typographic / layout) translation annotation (e.g. rating symbols, tags)
Usage Qualifier (2)	Must include specified content: URI of resource, with resource type (to include “credit”, “link”, “icon”, “text”)
Usage Qualifier (3)	Context of presentation: embedded in user-generated content embedded in original content
Usage Qualifier (4)	Time period: until recrawled permanently (eg for national library web archive) to date from date from date to date for a specified period from a specified event
Dependencies	“Crawl” and “Index” must be permitted. “Store” must be permitted for any Used Resource other than the original content item on the publisher’s website, or a link to the original item.

## ACAP Technical Framework – Public draft

Usage	Present
Note (1)	The word “present” has been chosen in preference to “display” or other possible alternatives because it is more suitable when the resource involves audio or other material that is not exclusively visual.
Note (2)	Highlighting of matched search terms in the display of a content item is specifically not regarded as constituting “modification” of the content or format.

**6. Other**

<b>Usage</b>	<b>Other</b>
Definition	Any usage not expressly permitted
User	Any
Used Resource	Any harvested and/or derived content
Used Resource Qualifiers	None
Usage Qualifiers	None
Note	An “Other” usage type is included solely in order to make an explicit statement that “any usage not expressly permitted is prohibited”. The corollary of including such a statement is that it must be possible to express all permitted usages fully and exactly.

**7. Glossary**

Content item	A piece of content on a publisher’s website, at any level of granularity which it is possible to reference in an ACAP permission statement.
Content item index entries	The <a href="#">index entries</a> (see below) required to make a content item searchable in accordance with a service’s indexing practices.
Content item link	A link to a content item on the publisher’s website.
Content item retrieved copy	A copy of a content item that has been retrieved by a crawler for processing. NB a retrieved copy may or may not be stored after it has been processed.
Content item snippet	A short indication of the content of a content item and, where applicable, its relevance to an end user search, normally generated by a service using proprietary algorithms.
Content item preserved copy	A copy of a content item that is preserved for use in a service.
Content item thumbnail	A thumbnail image of a content item webpage.
Content property	A property of a content item, encoded as part of the item or in associated metadata.
Crawler	Software controlled by a service operator that systematically retrieves content items from the web.
Index entry	An entry in an index that comprises (a) a look-up string derived from a content item retrieved copy, (b) a link to the content item, and (c) optionally, positional information to be used only for the purpose of increasing the effectiveness of searching and relevance assessment, eg by enabling the proximity of words in text to be taken into account.
Process	Any process that is operated by or for the operator of a crawler.
Publisher	A person or organisation that places content on the web.
Publisher content item summary	A summary of a content item supplied by its publisher.
Service	A service, including a search service, that provides organised access to content that has been harvested from the web.