# A Policy-based Institutional Web Archiving System with Adjustable Exposure of Archived Resources

Wasuke Hiiragi
Graduate School of Library, Information and Media Studies, University of Tsukuba
1-2, Kasuga, Tsukuba, Ibaraki 305-8550, Japan
ragi@slis.tsukuba.ac.jp

Tetsuo Sakaguchi
Graduate School of Library, Information and Media Studies, University of Tsukuba
1-2, Kasuga, Tsukuba, Ibaraki 305-8550, Japan
saka@slis.tsukuba.ac.jp

Shigeo Sugimoto
Graduate School of Library, Information and Media Studies, University of Tsukuba
1-2, Kasuga, Tsukuba, Ibaraki 305-8550, Japan
sugimoto@slis.tsukuba.ac.jp

## ABSTRACT
Despite the recognition that Archiving Web content is important and Web archiving systems freely crawl and collect resources on the Web for preservation, they have difficulties in collecting all versions of a single Web page and in preserving a collected resource with policies of use given by its creator. In order to solve the problems, we have proposed an Institutional Web Archiving System, which collects resources with the archiving policy given to the resources by its creator/provider, and we have proposed a set of consistency management functions for the institutional Web archives.

Based on our experience with the institutional Web archiving system (IWAS), we discuss usage restrictions of a resource archived in the IWAS and propose a technology to control browsing of restricted resources in accordance with the archiving policy given to the resource, e.g. access restriction to private information in a resource. We describe a function to hide protected portions of a resource using policy descriptions in the IWAS.

## Keywords
Institutional Web Archiving System, consistency of archived data, intranet, blotting out characters with black ink (called redaction).

## 1. INTRODUCTION
Organizations such as universities, governments, and companies provide information resources on the Web for public access. In addition, they provide resources for their members via their institutional Web. While most organizations make significant efforts to maintain their resources on the public and institutional Webs, archiving the Web resources is recognized as a crucial task to provide users with legacy resources as well as new resources.

There are global and nation-wide Web archiving services, e.g. Internet Archive in cooperation with the Library of Congress, the PANDORA archive in Australia, and WARP by the National Diet Library, Japan. These services collect resources through the global Internet, archive the resources based on their own policies, and provide them for use. The MINERVA project by the Library of Congress selects resources from archived Web pages and organizes collections based on specific topics, e.g. presidential elections, major incidents, etc.[1][2][3][4]

In general, Web archives are designed to provide archived resources in their original form and their original look-and-feel. Sometimes it is not possible to present an archived resource as it was not because of technological reasons but for content reasons, such as illegal content, privacy violations, and so forth. The conditions of use of archived resources may change over time and according to the environment of use. Sometimes, we are prevented from viewing a whole page because a small portion of it is not appropriate for display to the public. This problem could be overcome by simply hiding or modifying the offending part of the content.

We often see Web pages in which content is dynamically adjusted and displayed in accordance with a user's browsing environment. This technology seems useful to solve the problem mentioned above, i.e. hiding problematic portions on the fly. However, we have to cope with the copyright issues to revise the page for public browsing in accordance with the social environment at the time of use which could be 50 years from the date of publication of the page. This means that we need a description scheme for the conditions of use of archived resources.

In this paper, we propose a simple description scheme for the conditions of use of archived resources in an institutional Web archiving system, which we call IWAS. Institutional Web archives are slightly different from public Internet archives in terms of the coverage of the resources and collaborative working environment involving the content providers and archiving agents.

## 2. Issues in Using Archived Resources
### 2.1 Using Archived Resources
The primary goal of Web archive services is to prevent "born-digital" materials from disappearing in the future and to provide (open) access to archived collections. Thus, the primary use of archived resources is to view or download the archived resources.

A very basic issue at this point is that the provider of the archived resources is usually not the person or organization who created/published the archived resources. This means that from the stand point of the original creator/publisher, the use of their resources via an archive service is a secondary use of the resources. The archive service providers cannot take responsibility for the content of the resources but only for decisions making the resources available for their users. In other words, the archive service providers can choose to provide all or nothing of a resource even if only a small part of the resource is not

appropriate to be provided as it is. This inflexibility in providing archived resources is a fundamental issue for Web archives. It is not a technological problem but rather a management and policy problem. However, we need an appropriate technology to help managers of Web archives cope with this problem.

The basic framework to cope with the problem consists of the following points.

1. Identification of a rights management policy of the resource: If it is prohibited to revise a display or downloadable image of the resource, there is no way to solve the problem. We can just not display or download the resource from the archive.
2. Identification of the content of the resource that needs revision or removal: We need to identify the reason for the revision request and conditions to revise the resources. The conditions of use of a resource are not independent of thesocial environment of the resource. A typical example is a request to hide personal names for privacy reasons.
3. Identification of location of the content that needs revision or removal: We need to identify the location of problematic content in a resource in order to revise or remove the content. Automatic identification of the location of problematic content is, in general, challenging. Technologies such as topic search and person identification are crucial. Hybrid approaches that combine those technologies and heuristics would be essential.
4. Revision or removal of the content from the resource presented to users: The portion which cannot be shown to the users has to be removed from the source or revised to a format which satisfies the requirements. The former means that the target portion is removed from the original resource but the latter means the portion can be kept in the source but hidden from the audience.

In this paper, we mainly discuss a technology to hide the portions of a resource that need to be hidden from the audience in the context of institutional archives. Revisions by an archive should be easily identifiable and, more importantly, the revisions need to be restorable. Institutional Web archives are significantly different from Internet Web archives in the following two aspects,

1. collaborative collection and maintenance between content providers and archive managers, and
2. the high possibility of collecting resources which are to be hidden from the general public.

Content providers can give archival managers a policy of maintenance and accessibility from the public from a viewpoint of long-term use; some resources or some portions of the resources need to be hidden from the public for a designated period of time, such as 30 years. For example, archives blot out a portion of a document when they allow public access to the document .

The next section overviews an Institutional Web Archiving System (IWAS) which the authors developed [5][6]. Section 3 describes a technology to embed codes to aid access to the resources with revision on-the-fly.

## 2.2 Institutional Web Archiving System (IWAS)

### *2.2.1 A Concept of Institutional Web Archiving*

Organizations such as universities and governments have serious demands from their members to archive and preserve their Web resources. Although there are well-known Web archiving services, technologies for the organizations to archive their Web resources are still immature. This model assumes cooperation between resource providers and archiving agents. The fundamental difference between global Web archiving and organization-oriented Web archiving is the incentive of the organizations to archive their contents. The core concepts of this model are description of an archiving policy associated with every resource to be archived and the metadata used in storing the Web pages.

The Web archiving system proposed obtains a Web resource each time it is revised. This revision includes updates of the content of the resource, removal of the resource and relocation of the resource. In order to obtain all revisions of the resource, the archiving system has to be notified each time the resource is revised. The system has an interface for receiving events of revision from Web servers, which we call *update events*. The Web servers that cooperate with the archiving system determine archiving factors which are, for example, events to initiate archiving, a method to archive, access control of archived resources, and so on. These factors are associated with the resources as policy descriptions and stored in the Web servers. The policy descriptions are used by the Web servers and the archiving system.

An archived Web page is organized into a set of *Archived Components*, an *Archived Resource*, and a *Set of Archived Resources*. An *Archived Resource* is composed of one or more *Archived Components*. An *Archived Resource Set* is a series of resources collected from a single location or a single site which may have changed its location. And we call the object which added the *Access Policy* to these, the Web Archive Object (WAO). [Figure 2.1.]

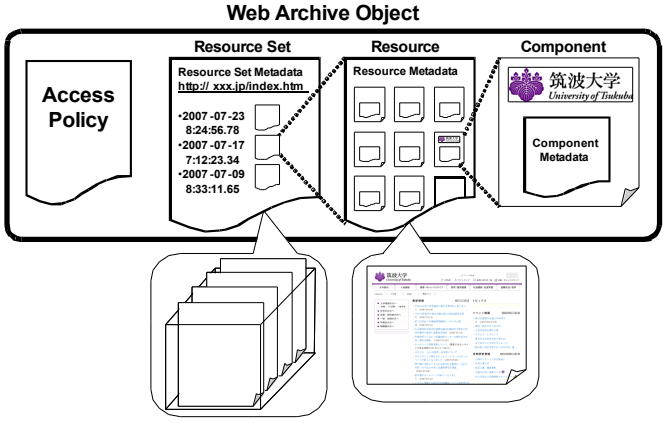


**Figure 2.1. Web Archive Object Model**

***Archived Component.***

An *Archived Component* is a primitive object that is stored in the archive. An archived component is a file stored in a location identified by a URL. It is identified in the system by its URL and date-and-time when it was captured. Components are used

to reconstruct an *Archived Resource* on a client, e.g. Web browsers, when the resource is requested.

***Archived Resource.***

An *Archived Resource* is a composite instance composed of one or more *Archived Components*. It is a snapshot of a Web page at the time the page was captured. An archived resource is identified by the location of the page and the date-and-time of capture.

***Archived Resource Set.***

An *Archived Resource Set* is an ordered set of *Archived Resources* captured from a single location or a logically single site. *Archived Resources* in an *Archived Resource Set* are ordered by date-and-time of capture. Every *Archived Resource Set* has a date-and-time of creation and that of removal. The creation and removal date-and-times are used to identify the period of existence of the *Archived Resource Set*.

***Access Policy.***

The archiving system has a set of access control functions to restrict access to archived resources in accordance with their archiving policies. Restrictions are determined by network range, user, date and time, and so on.

Metadata is created for each instance of an *Archived Component*, an *Archived Resource*, and an *Archived Resource Set*.

### *2.2.2 Web Archive Object and Collection Policy*

An IWAS has to have a collection policy (CP) which describes criteria and conditions to collect resources from an institutional Web. The IWAS may change its collection policy from time to time. Every WAO is associated with a CP which is effective at the time of creation of the WAO. As shown in Figure 2.2, every WAO is associated with a CP effective at the time of creation.

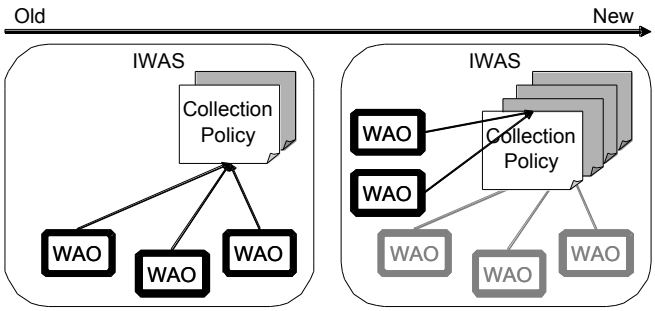


**Figure 2.2. Collection Policy and WAO**

### *2.2.3 Reorganization and Merge Policy*

Naturally, organizations change over time. This means that an IWAS has to cope with organizational changes. The policy descriptions are key components for the IWAS. However, the policy descriptions need to be changed along with an organizational change. In order to achieve this, we have included a Merge Policy (MP) description in [6]. An MP specifies a method for how metadata in a WAO is to be changed, and a list of *Archived Resources* and *Archived Components* which have to be deleted. Figure 2.3 shows a merger of old IWASs to a new IWAS.

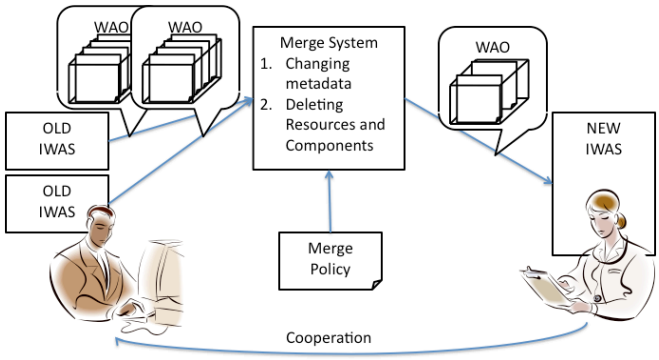


**Figure 2.3 Merge Policy and Merge Method**

## 3. Discussion

### 3.1 Discussion and Lessons Learned from the Institutional Web Archiving System

In these previous studies, we have clarified the following four functions for IWAS - Collecting, Archiving, Accessing, and Merging.

These functions work with policy descriptions, as the main feature of IWAS. These policy descriptions are collected and archived within a WAO. They are, in general, prepared by creators/providers of the resources. However, writing a policy description is a costly task for the creators/providers. In addition, the policy descriptions are not always good enough for the IWAS to perform its task properly.

Besides this, we think that there are problems in Collecting and Accessing functions.

*Collecting*

This function gathers and stores Web resources with metadata which is automatically created based on the CP. A WAO will not be created for a resource which does not match the policy. This function re-collects resources each time it is triggered to start collection based on the crawling policy of the CP.

It does not record the revisions – which part of a resource is revised, removed or added. It is desirable to record revisions in a resource in order to realize more flexible functions such as access control to a part of a resource but not to the entire resource.

*Accessing*

In the IWAS, access to WAOs is controlled by an Access Policy (AP) associated with the WAOs. Access control is carried out resource by resource but not by a portion of a resource. This means that it is difficult to implement a service to hide specific content parts from users, e.g. name of a person and telephone numbers. Masking of a specific portion in a document is frequently used by archives when they provide a document which includes privacy information.

### 3.2 Related Works

There are many projects and much research related to Web archiving and preservation.

The importance of Web archiving is obvious. International Internet Preservation Consortium (IIPC) has a mission to acquire, preserve and make accessible knowledge and information from the Internet for future generations everywhere. The IIPC's Preservation Work Group says "work on maintaining accessibility for the long term remains reasonably undeveloped"[7]. In its IP Strategic Program 2008, the secretariat of Intellectual Property Strategy Headquarters, Cabinet Secretariat in Japanese refers to the "Creation of the framework to promote content distribution in consideration of new ways of utilizing content"[8][9].

IIPC is conducting several projects. Among them, the Web crawler Heltrix and the metadata schemes are crucial components for the Web archiving community. However, they are primarily designed for an open Web environment and no cooperation between content providers and archives is assumed. In this respect, the study presented here is different from those conventional open Web archiving systems.

Preservation of dynamic content and/or its look-and-feel is a crucial issue for the Web archiving community but is hard to solve. Kwout and Web Gyotaku are undertaking archiving/preservation with screenshots [10][11]. This leaves unsolved the problem of how to hide specific parts from users when they browse archived resources.

# 4. Functional Requirements of IWAS

## 4.1 Expansion of Access Policy

In this paper, we propose a function of access control to an *Archived Component*. We propose a technology to hide a specific portion in a Web page based on the IWAS framework – policy descriptions determined by cooperation between content providers and archive managers.

We think the following three functions are required to hide information of Web pages:

1. Hiding words and phrases in an *Archived Resource* based on a policy of creator,
2. Extending the function 1. to apply it an *Archived Resource Set*, and
3. Recognizing character strings to be hidden when this is ordered by a policy.

The first function is the most basic requirement. Users should not be able to get the information that is to be hidden according to a policy given by the archive.

The second function simplifies the process to specify the words and phrases to be hidden. The hidden words and phrases may be included in more than one *Archived Resource* in a single *Archived Resource Set*. Users can specify the words/phrases on a single archived resource, then this function can hide the words/phrases in other archived resources in the same *archived resources set*.

The third function helps to specify words and phrases that should be hidden as well. An archived resource could include synonyms of a hidden word/phrase. Those synonyms should be hidden as well.

These functions are required to enhance the IWAS. The current implementation, however, includes only the function 1.

## 4.2 Basic concept for Adjustable Exposure on Archived Resources

In the IWAS environment, not all of the resources are intended to be open to the public, which is a significantly different from in an open Web archiving system. For example, a Web page which includes personal information, e.g. a personal name and email address, should be restricted for use based on the access policy given by the content provider and archive manager regardless of time since the creation of the resource.

The fundamental issue in access control is to make available a part of or whole of a resource in accordance with the access policy over time. As a resource or a component of a resource which does not conform to the collection and archiving policy of the IWAS is excluded in the archive, all of the resources in the archive, in principle, conform to the policy at the time of collection. Over time, the policy could be changed, so that the IWAS needs to organize the archive in accordance with the policy, as briefly mentioned in section 2.2.3.

An access policy to determine the conditions of use in accordance with the users and the environment of use can be determined independently of the collection and archiving policy, e.g. browsing-only or downloadable as a file, entirely or partly browseable, and so forth. Thus, the Access Policy mainly determines the accessible portion of an archived resource and the format and method of access to the resources.

In this paper, we discuss a function to determine how an archived resource should be presented to users. We call this function Adjustable Exposure of an archived resource in this paper. The function only adjusts the visibility of the content in the archived resource and does not change any content of the resource.

The technique used in this paper for adjusting exposure of a resource is masking content to be hidden from users in accordance with the access policy. Because the masking function looks like a blotting out of the characters/words with black ink, we call the proposed technology Blotting Out Characters (BOC). In this information hiding method, we hide a portion of a resource as shown in Figure 4.1, e.g. character texts, words, phrases and so forth.

Masking characters in a document is a conventional function. The advantage of the BOC proposed in this paper is that the portion to be masked is determined by the policy descriptions and the structural information of the archived resource and that of the archived resource set. Thus, the BOC is not a simple masking of a single page but a structural masking based on the structure of an archived resource set and the policies.

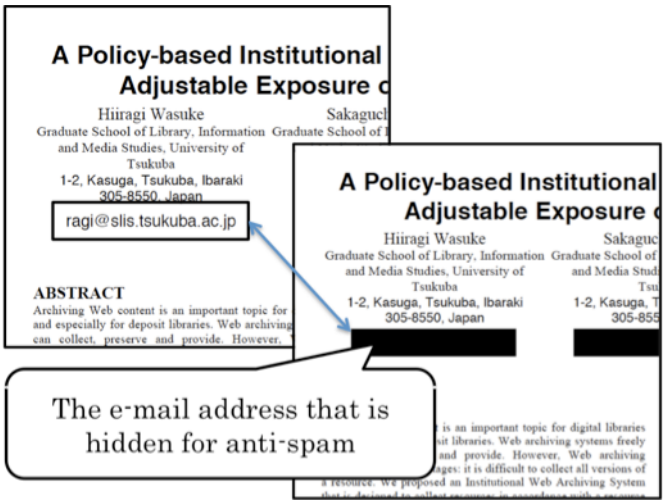


**Figure 4.1. Example of Blotting Out the Characters**

# 5. Blotting Out Characters System Model

## 5.1 Overview

The BOC system has the following three functions,

(1) Setting the BOC Region of an *Archived Resource*,
(2) Revising the BOC region and the conditions, and
(3) Providing the Archived Web page with some content hidden.

The first function is for a creator of a resource to specify a region(s) where characters and words are to be hidden in accordance with the access policy and the conditions of use.

This function is the most basic function of the Blotting Out Characters (BOC) method discussed in this paper. Because a blotting is left on a published Web page, the user understands that there was content that a creator wanted to hide in the Web page. In addition, the user can suppose what kind of information there was with the context of before and after. And, the user understands that some words are hidden according to a policy if the creator gives a reason for the BOC as metadata.

The second function revises a BOC region and conditions associated with the region. IWAS has to cope with changes in various aspects – organizational change, collection policy change, archival policy change, access policy change, and scheduled and non-scheduled service change along with the archival process. The second function is required in order to revise the regions in accordance with the changes.

We designed the BOC method to implement these functions. Figure 5.1 shows the flow of the IWAS to set a BOC Region and Condition Description into a main body.

(1) A creator writes some BOC Regions into a Web resource which is collected as an Archived Resource,

(2) The creator makes a Condition Description and attaches it to a *Archived Resource Set*,

(3) A BOC System gets the newest revision of conditions and an *Archived Resource* from the *Archived Resource Set*, and

(4) The BOC System hides words based on (1) and (2) in the Archived Resource, and provides it to the requested user.

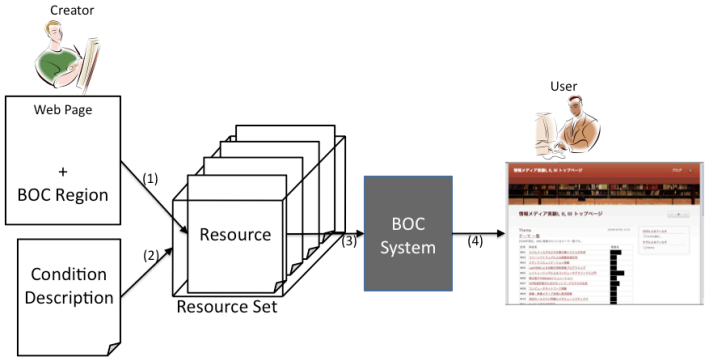


**Figure 5.1. Structure of BOC method**

The BOC system manage the specification of the BOC region and the decisions about hidden words, changes of hidden words in every BOC region, and the provision of the Web page when it is finished with the BOC processing. In the next section, we explain each function.

## 5.2 Setting the BOC Region

A creator specifies a BOC region of a resource which he/she thinks is necessary in the archival process in future. The creator specifies the BOC region using XHTML tags. He/she assigns an identifier (ID) to the BOC region. The ID is the character string that is unique in each *Archived Resource Set*. The creator puts the ID in words/phrases that the creator thinks have the same meaning. This element is written in RDF, and the structure is shown in Figure 5.2.

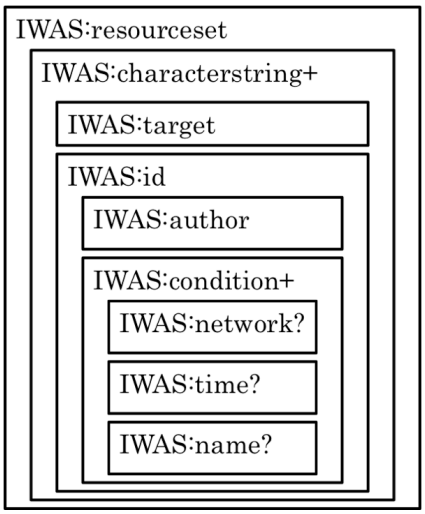


**Figure 5.2. BOC Schema List**

By Figure 5.2, the element with "?" shows an optional element, with "+" to show how many times.

These elements have relations as in Figure 5.3 when they are expressed in an RDF graph. In the Figure, the upper part expresses a BOC Region of a Web resource which is collected as an *Archived Resource*; the bottom part expresses a condition of the BOC Region. These are related by IWAS:ID.

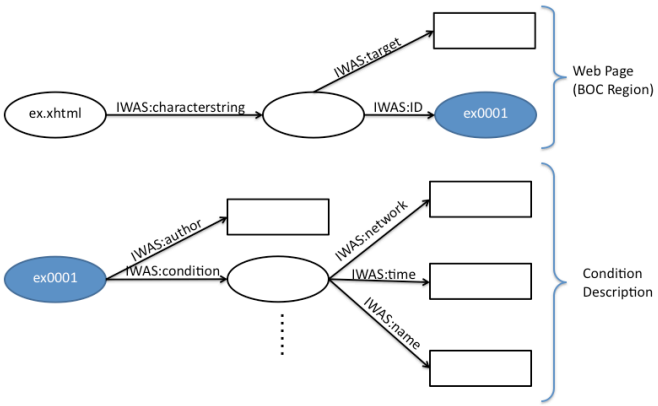


**Figure 5.3. BOC Schema in RDF Graph**

This system merges two RDF graphs with the node of IWAS:ID. The metadata to be used by the BOC system gives conditions for when to hide the relevant content portion.

## 5.3 Revising Conditions

The revision condition is managed for each *Archived Resource Set*. As time goes by, the conditions of BOC may change for various reasons. The condition part of BOC is separate from the *Archived Resource*. Therefore, resource creators/providers and archival administrators co-edit the condition part. The system can maintain consistency by keeping track of the change history of a resource set.

The BOC Region reduces the cost of the creator's task to write a Condition Description. Even if a uploaded Web page with a BOC Region is collected as an *Archived Resource*, a Condition Description is related to a lot of BOC Regions by IWAS:ID as in Figure 5.4. Therefore, even if the number of words/phrases to be hidden increase and decrease, the BOC system can hide the words/phrases.

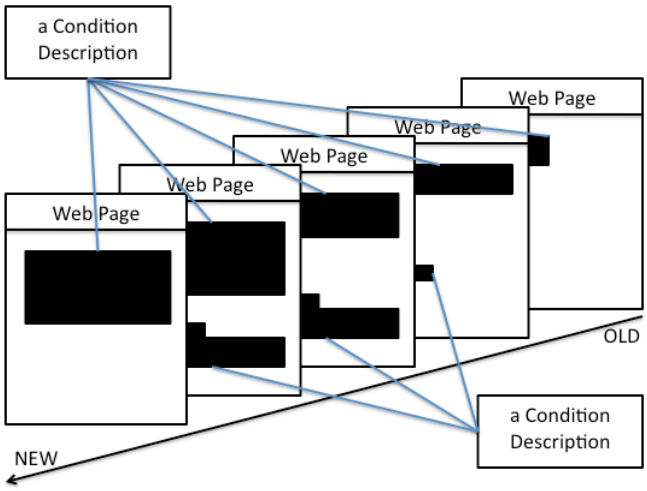


**Figure 5.4. Relating BOC Regions to a Condition Description**

## 5.4 Providing Archived Web Page

With a main body which is a publication object and BOC region, and the, The BOC system, as a publication objct, has a BOC region and descriptions of the condition for hiding content.. It has three kinds of attributes: IWAS:name, IWAS:time, IWAS,network, which are used to specify user conditions. The meaning of the attributes is as follows:

- IWAS:network
  - Which network does a user request from?
- IWAS:time
  - When does a user request it?
- IWAS:name
  - Which user requests it?

The BOC region is hidden by a system if a user does not satisfy the conditions. Figure 5.5 shows the flow of the data in the IWAS to provide an archived resource in an *Archived Resource Set*.

(1) A user requests the IWAS to show an archived Web page with his/her metadata,

(2) The BOC System in the IWAS finds an *Archived Resource* from an *Archived Resource Set*. A description of the BOC is included in this *Archived Resource*,

(3) The BOC System gets the newest revision of conditions from the *Archived Resource Set*. And the system checks the user's metadata, the *Archived Resource,* and the conditions, and

(4) The BOC system hides words and provides a Masked Web page to the requesting user.

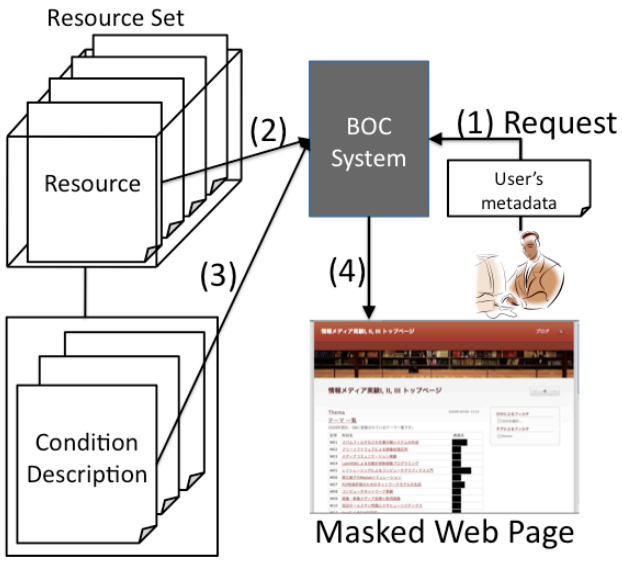


**Figure 5.5. Providing Part of BOC System**

# 6. Implementation

The system consists of three descriptions which are XHTML descriptions with the BOC region described in an *Archived Resource*, Conditions Description, and User's Metadata. The BOC system interprets them. In addition, we build this structure on top of the IWAS.

A BOC region is described in RDFa/XHTML. Figures 6.1 and 6.2 show an sample description and its RDF graph, respectively.

```
<html
 xmlns="http://www.w3.org/1999/xhtml"
 xmlns:IWAS="http://jungle.slis.tsukuba.ac.jp/"
 xmlns:IWASID=”http://jungle.slis.tsukuba.ac.jp/cond1.xml”
>
<head><title>My Email Address!</title></head>
<body  about="examplepage.xhtml">
My name is Wasuke Hiiragi. My e-mail address is
<span rel="IWAS:characterstring"><span property="IWAS:target
">ragi@slis.tsukuba.ac.jp</span><link rel="IWAS:ID" resource=
"IWASID:ex0001"/></span>.
</body>
</html>
```

**Figure 6.1. Example of XHTML Description with BOC Region**

The RDF graph in Figure 6.2 is created by matching the description above with a description of masking conditions. The Blotting out function is performed on the Web page using this graph.

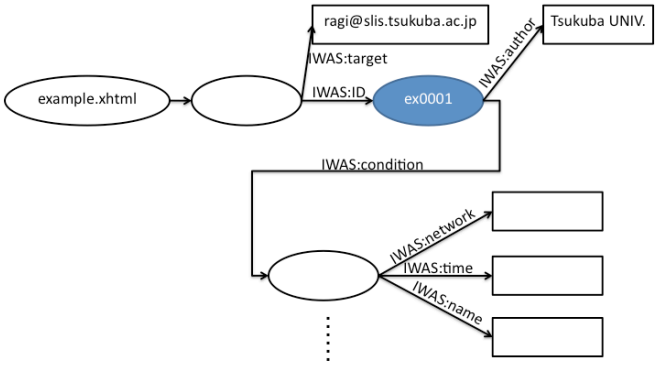


**Figure 6.2. Example of BOC Region RDF Graph**

## 7. Discussion and Conclusion

This study aims to promote the use of archived Web resources based on an institutional Web archiving environment.

A creator/provider cannot know the date when archived resources will be used in the future. Therefore, archived resources are not always browseable by everyone as they were stored in the archive because of non-technological issues such as privacy and copyright. We think that it is not reasonable to hide an entire resource which includes a portion which cannot be open and that archives should have functions to provide as much as possible of the resource. This issue is not be typical in conventional Web archiving systems because they collect open resources but is crucial for institutional Web archives.

However, we could not implement a function 2 and 3 of Section 4. So we want to implement the functions based on finding "typical update work", and methods of estimating content which is changed on the basis of generations of Archived Resources.

Archiving Web content is an important topic for digital libraries and especially for deposit libraries. We believe our ideas will be important in the future management of archived Web pages.

## 8. ACKNOWLEDGMENTS

The authors would like to express our thanks to Profs. Atsuyuki Morishima and Mitsuharu Nagamori for their useful discussions on this study. This research is supported in part by Grant-in-Aid for Scientific Research (B) #19300081.

## 9. REFERENCES

[1] Internet Archive
http://www.archive.org/ [accessed 2009-07-23].

[2] PANDORA
http://pandora.nla.gov.au/ [accessed 2009-07-23].

[3] WARP
http://warp.ndl.go.jp/[accessed 2009-7-23].

[4] MINERVA
http://lcweb2.loc.gov/diglib/lcwa/html/lcwa-home.html
[accessed 2009-07-23].

[5] Wasuke Hiiragi, Tetsuo Sakaguchi, Shigeo Sugimoto, Koichi Tabata: “A Policy-based System for Institutional Web Archiving”, Z.Chen et al.(Eds):ICADL 2004,LNCS 3334, pp.144-154,2004.

[6] Wasuke Hiiragi, Tetsuo Sakaguchi, Shigeo Sugimoto: “An Architecture of Institutional Web Archiving System that have functions to Merge and Split Archives” Japan Society of Information and Knowledge, 2008. in Japanese.

[7] IIPC
http://netpreserve.org/about/index.php [accessed 2009-07-23].

[8] Secretariat of Intellectual Property Strategy Headquarters
http://www.ipr.go.jp/ [accessed 2009-07-23].

[9] IP Strategic Program 2008 [accessed 2009-07-23].
http://www.kantei.go.jp/jp/singi/titeki2/keikaku2008_e.pdf [accessed 2009-07-23].

[10] Kwout
http://kwout.com/ [accessed 2009-07-23].

[11] Web Gyotaku
http://megalodon.jp/ [accessed 2009-07-23]