Metadata and PICS Management

William Song

SISU, Electrum 212, S-164 40 Kista, Sweden

Abstract

Almost infinite resource space in the Web provides massive information for various purposes. Problems arise: How to effectively obtain the Web information that we need, and how to block the inappropriate Web materials. PICS provides a standard for possible support to the Internet users to control access to the Web. To a broader sense, METADATA is introduced to the Web information structure and gives a more powerful management in modelling the information. In this report, we study the current situations of PICS development, survey the existing work, and focus on the access control of the Internet information and metadata related issues. We consider our future effort will be put in the metadata analysis and formulation.

Contents

1 Introduction

Rapid development of Internet and related facilities have made Internet a major information resource for Web information readers. Billions of bytes of information are poured into and fetched from the Internet. To select appropriate Internet content is becoming increasingly needed by Web readers. The need for selection concerns two aspects: controlling accessing and guided investigation to the Web information.

Substantial information is obtained from various sources in the Internet. However, on one hand, not all information is suitable for readers of all walks. For instance, parents may not like their children to be exposed to movies and pictures with harmful materials from the Internet. They may hope to set a sort of restrictions (filters) on their children's access to the Internet. That is, it is required to provide facilities for controlling the Internet access. On the other hand, substantial information makes it difficult to find out what an Internet reader is really interested in and expects to obtain. Therefore, to guide readers to search (i.e. guided search) from the Internet what they want becomes an important issue.

These two aspects make up the main subjects in the PICS (standing for Platform for the Internet Content Selection) area. More related issues include electronic publishing – where a publisher may give different readers different rights to access an Internet-published book, copyright protection – how an electronically formed property is protected from embezzlement, netnews bulletin board – where various readers and posters expect to be directed to subjects of interest to him or her.

Recently, a more general concept, "metadata", is introduced in PICS to describe the Internet data. Content labels, that are used in PICS for rating the Web documents, can be viewed as a set of attributes associated a metadata. That is, metadata uses a list of attributes or labels to characterise a Web document in order to categorise the Web information items (or documents) for information search and access blocking. Compared with its narrow sense, where it mainly describes rating systems and rating services, now PICS is somewhat considered to cope with the problem of the Web document structuring and categorisation. We will give a detailed discussion on the metadata concept later on and focus on the metadata research in our future work.

1.1 What is PICS?

PICS is a technical platform that offers a highly flexible tool for filtering Internet content. It does not rate the content but empowers any individual or organisation to develop their own rating systems, distribute labels for Internet content and create standard label-reading software and services. PICS gives the user customised access to Internet content.

"PICS is a major step forward in the evolution of the Web and is another example of how the W3C is working to make the Web easier to navigate," said Tim Berners-Lee, Director of the World Wide Web Consortium and creator of the World Wide Web. "PICS will allow Web users to find information they want and avoid information they would prefer not to see. PICS is also beneficial in a wide variety of applications ranging from security and privacy protection to searching digital libraries."

"PICS establishes Internet conventions for label formats and distribution methods, without dictating a labelling vocabulary," said Jim Miller, of the W3C, Co-Chair of the PICS Technical Committee. "It is analogous to specifying where on a package a label should appear, and in what font it should be printed, without specifying what it should say."

1.2 A general view of PICS applications

PICS is considered to be a platform for structuring the Web information for certain content selection, so that the Web information can be accessed according to rules of fetching information items needed and blocking inappropriate information items.

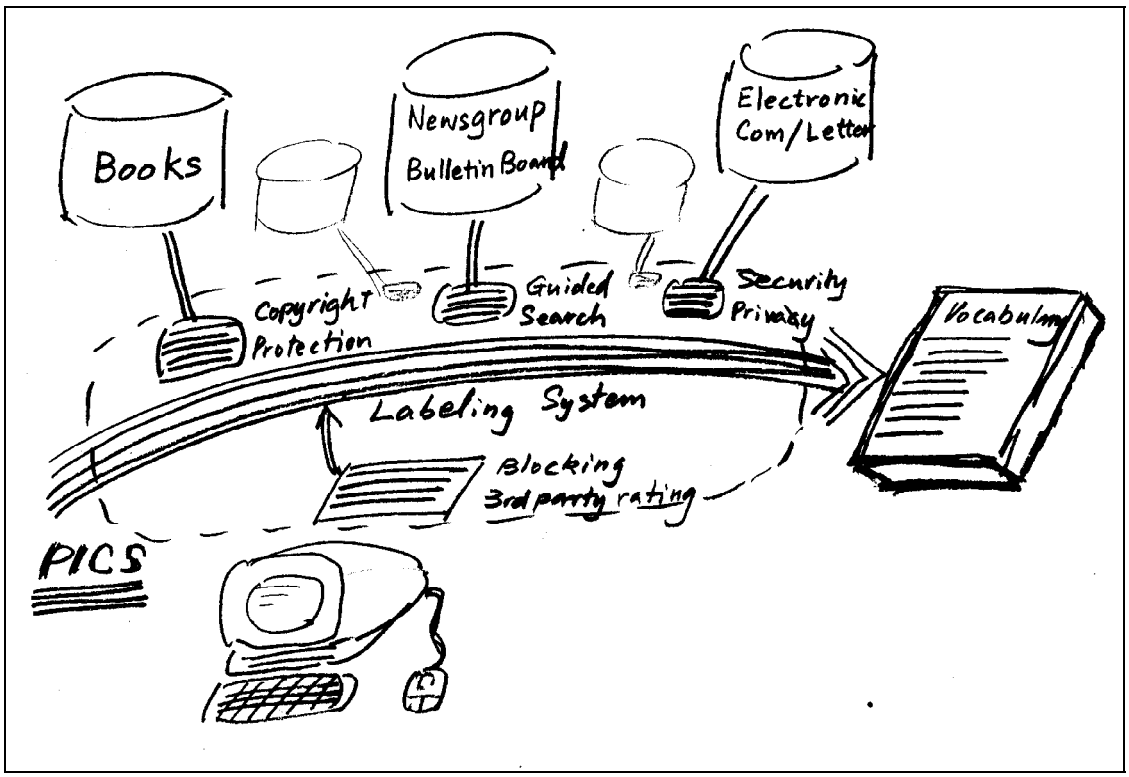

Fig 1. PICS supports labelling systems to access to Web resources.

Examples of applications in PICS are, among others, copyright protection when accessing to publications, news groups and bulletin boards when people are grouped in terms of their interests, and digital signature when access security is needed for electronic commercial and private communication, see Fig 1.

Control of access to the Web information can be provided based on a rating system and a vocabulary, where semantics of information contents can be described in a sort of metadata definition.

Once a suitable metadata framework is defined, the Web information items can be searched by the Web readers through an information-guided facility and blocked when they are considered to be inappropriate for recipients.

1.3 Social Contexts

Obviously, the vast Web resource space, on one hand, provides a great deal of information to empower people in gaining a broader knowledge of any kind, on the other hand, contains illegal, harmful and inappropriate materials for certain groups of people. Consequently, the access to and use of the Web information has received attention from a variety of organisations, such as governments. In this section, we discuss the opinions from the social contexts.

1.3.1 Public Education

Strong influences of the massive Web information has increasingly reached public education, such as schools, not to mention universities. The Web provides materials far more beyond textbooks and with all dimensions of subjects in schools. Richness, attractiveness and effectiveness in the Web materials enable school education to be more diverse and multi-cultural. It helps not only pupils and students to be able to learn (perhaps by themselves) intensive knowledge from accessing the Web, but teachers to understand more contextual knowledge in what they teach as well.

However, attention has been paid by experts to the unsuitable Web information being exposed to pupils and students of different ages. Information blocking techniques for the Web information is introduced in public education for protecting minors from being influenced by harmful Web materials [Etiken97]. The blocking of inappropriate materials depends basically on three factors: 1) The supervisor – Parenting styles differ, as do philosophies of management and government; 2) The recipient – What's appropriate for one 15-year-old may not be so for an 8-year-old; and 3) The context – A game or chat room that is appropriate to access at home may not be appropriate at work or school.

An approach of third party rating provides a possibility of blocking harmful materials from the Web by parents and supervisors. A rating system, along with a vocabulary, can support parents to select suitable information for their kids. Moreover, the system can, based on different levels of label settings, offer different access levels to recipients according to what parents set in the system [Resnick 96-CACM].

1.3.2 Government Interest

An important argument regarding rating the net is about the net information which is harmful or illegal. The European Commission approved a *Communication on harmful and illegal content on Internet* and a *Green Paper on the protection of minors and other humans* in the context of new electronic services.

While the Communication gives policy options for immediate action to fight against harmful and illegal content and concentrates on the Internet, the Green Paper takes a horizontal approach and will initiate a medium- and long-term reflection on the issue across all electronic media. Both documents advocate a closer co-operation between

Member States and, on an international level, the use of filtering software and rating systems, and an encouragement to self-regulation of access-providers.

EU concerns

Driven by its meteoric growth, and rapid evolution from a government/academic network to a broad-based communication and trading platform, the Internet is currently revolutionising a number of economic sectors, with the emergence of a vibrant and fast-growing "Internet Economy". Simultaneously, the Internet has also become a powerful influence in the social, educational and cultural fields – empowering citizens including educators, lowering the barriers to the creation and distribution of content, offering universal access to ever richer sources of digital information.

Reflecting these opportunities, the vast majority of Internet content is for purposes of information for totally legitimate (and often highly productive) business or private usage. However, like any other communication technologies, particularly in the initial stages of their development, the Internet carries an amount of potentially harmful or illegal contents or can be misused as a vehicle for criminal activities. Although statistically a limited phenomenon, a wide range of distinct areas are concerned. These are covered by different legal regimes and instruments at the national and international level, such as national security, protection of minors, etc.

While the benefits of the Internet far outweigh its negative aspects, these aspects cannot be ignored. They are pressing issues of public, political, commercial and legal interest. Reflecting these concerns, recent political discussions in the European Union have stressed the need for urgent action and concrete solutions.

Therefore, on the 27th of September, 1996, the Telecommunications Council adopted a resolution on preventing the dissemination of illegal content on the Internet, in particular child pornography. Stressing the need for rapid response, the Council urged the Commission to carry on its ongoing work and to present practical measures in time for the next Telecommunications Council. In response to this, for example, the Swedish government has issued a Green Paper to Swedish organisations, administrations, and companies for consideration of suitable actions [KU97].

The Commission is fully aware of the importance of these issues, and of the need to strike the right balance between ensuring the free flow of information and guaranteeing protection of the public interest so as to meet justified concerns.

Social impacts – harmful and illegal materials

Illegal content is primarily a matter for national authorities (the judicial system) and the nature of Internet may constitute severe problems in enforcing national laws against materials from other countries where the materials are perhaps legal. Legislation, dealing with materials which are illegal when distributed to young people but not for adults, may vary from country to country.

Apart from the political arena – in legislation and at both national and international levels, the issue of harmful and illegal content on Internet has been addressed also in other contexts. It is one of the most frequently discussed issues, in media like newspapers, magazines, television etc., about the appropriate use of Internet. These

issues are also discussed within parental organisations, youth and school authorities and other organisations concerned with the well-being of minors.

Efforts have been made to provide users with tools to control access to specific kind of information. Software packages are also available to support access to beforehand-approved sites and/or block access to other inappropriate sites. This approach requires that people (parents, schools, software vendors) actually have control of suitable/ unsuitable sites. An ISP can also offer services of this kind.

In several environments, like in schools and at home, adults would hope (and take responsibility for) that the access to Internet gives children opportunity to find relevant and useful information instead of being exposed to harmful materials. Determination of both what is useful and what is harmful depends on cultural values, context, age and varies from time to time. Governmental initiatives in US and other countries has been issued to provide searchable information about the content of governmental agencies information archives (GILS).

Labelling support in PICS provides a capability to rate and consequently to filter the net content so that useful materials can be accessed and harmful ones avoided.

Australian concerns

In a working paper from The Australian Broadcasting Authority (ABA), it can be summarised that community expectations, industry concerns and user interests can be best addressed:

- by the implementation of a voluntary two tiered approach to codes of practice;
- by recognising that minors should be protected from harmful material and introducing accreditation standards and quality control processes for services providing managed 'safe' areas for minors and cautious adult consumers;
- in the consistent application of standards across on-line services as reflected in the appropriate sections of the National Classification Code and the Classification Guidelines;
- in the provision of consumer advice about the content standards applied and the actual or likely content of services and products being provided on-line, with more detailed labelling and classification requirements being applied to accredited services;
- by an easily accessible, transparent, quick response and quality controlled complaints handling process;
- by an independent and impartial assessment body with expertise in the application of community standards in the assessment or classification of media content;
- in the enforcement of legislative provisions which require compliance with content standards set out in the codes of practice and the determinations of the independent assessment body;
- by an on-going research effort addressing a variety of issues relating to the use, content, regulation and provision of on-line services; and
- by an extensive education campaign targeting consumers, parents and others involved in the care of children and young people, users and service providers.

1.4 Overview

As stated above, the vast Web information gives rise to both opportunities and problems with respect to the controls of access to and use of the Web materials. PICS proposes a standard on content labelling to support such controls. In this report, we attempt to discuss the current PICS situations in the following three aspects:

1) PICS specifications include, in a narrow sense, rating systems, rating services, and a vocabulary provided either by the information source providers or the Web readers. The specifications suggest a standard to follow to design one's own filtering mechanism. We shall describe the specifications in the next chapter. Some related work out from PICS, will also be discussed in the chapter, such as filtering, metadata, and so on.

2) More concretely, PICS issues concerning a platform for content labelling, the Internet access control, and metadata transformation and formats will be depicted in chapter 3.

3) In chapter 4, we focus on some PICS application scenarios and codes, which may help the readers to better understand what PICS attempts to achieve and how the Web access control is carried on.

We conclude the report by pointing out our aim to continue research work regarding PICS issues. We will mainly focus on the metadata extraction and the content label modelling. The integration issue will be introduced to support the metadata (labels) schema modelling and integration.

2 PICS Work

2.1 PICS specifications

This section describes the PICS specifications, including rating services, rating systems, and content labelling. The original PICS technical specifications are available in [Miller96]. We use this description of PICS specifications with small modifications.

2.1.1 Rating Services

A rating service is an individual, a group, an organisation, or a company that provides content labels for information on the Internet. The labels it provides are based on a rating system (see below). Each rating service must describe itself using a newly created MIME type, application/PICS-service. Selection software that relies on ratings from a PICS rating service can first load the application/PICS-service description. This description allows the software to tailor its user interface to reflect the details of a particular rating service, rather than providing a "one design fits all rating services" interface.

This specification does not state how the application/PICS-service description of a rating service is initially located. For users of the World Wide Web, it is expected that well-known sites will provide lists of rating services along with their application/PICS-service descriptions. It is also expected that client programs will cache copies of application/PICS-service descriptions, so any incompatible change in a service description should be accomplished by creating an entirely new service URL.

Each rating service picks a URL as its unique identifier. It is included in all content labels the service produces, to identify their source. In general this identifier includes a version number to simplify transitions due to incompatible changes over time. For example, a sample service "http://www.gcf.org/v1.0/" includes "v1.0" as its own version number. To ensure that no other service uses the same identifier, it must be a valid URL. In addition, the URL (when used within a query) serves as a default location for a label bureau that dispenses this service's labels (see PICS Label Distribution at http://www.w3.org/pub/WWW/PICS/REC-PICS-labels-961031.html).

Since the service identifier is a URL, it can be used to retrieve a document. That document may be in any format, but it is recommended that it:

- is in HTML format;
- is organised for ease of use by novice computer users (the application/PICS-service description would be a poor choice);
- describes not only the rating service, but the rating system as well (or provides a link to another document describing the rating system);
- is available in multiple languages, either through an existing negotiation mechanism or through links to alternate language versions.

2.1.2 What is a "rating system"?

A rating system specifies the dimensions used for labelling, the scale of allowable values on each dimension, and a description of the criteria used in assigning values. For example, the MPAA rates movies in the USA based on a single dimension with allowable values G, PG, PG-13, R, and NC-17.

Each rating system is identified by a valid URL. This enables several services to use the same rating system and to refer to it by its identifier. The URL naming a rating system can be accessed to obtain a human-readable description of the rating system. The format of that description is not specified.

2.1.3 What is a "content label"?

A content label (or rating) contains information about a document. As described in PICS Label Distribution, a content label (or rating) has three parts:

1 the URL naming the rating service that produced the label;

2 a set of PICS-defined (and extensible) attribute-value pairs, which provide information about the rating, such as the date when the rating was assigned;

3 a set of rating-system-defined attribute-value pairs, which actually rate the item along various dimensions (also called categories).

2.2 Filtering

Information filtering is viewed as a mechanism for the Web readers to get from the Web what they need and abandon what they don't. Content-based filtering is a basic approach to filtering the Web information. The approach uses keywords and/or attributes of contents, usually in information items (such as an article), obtained from the Web sources to form a basic information content for a parsing pattern, which the users define by providing their own keywords and attributes for the information they like to obtain. The information from the Web will pass through the pattern if their keywords and attributes fit the user’s requirements.

Generally, keywords or attributes to describe a Web document are considered to be labels in PICS. In its application programs, PICS specifies a set of labels which describe the information of a certain group of documents. When accessing to the Web sites, these documents can be filtered (either allowing to obtain or blocking) by the programs with label set. Furthermore, different scales of values can be assigned to the labels (attributes) to give a more specific assessment of the materials in the Web. For instance, a (PICS rule) statement {Filter (Block "Materials.Graphics > 3)} will allow the Web readers to obtain materials with less graphics (rating from 0 to 5).

The limitations of this basic approach are that 1) information items should be of some machine parsable form (it would not be parsable if information items are in sound, video, etc); 2) the filtering results contain many replications that the users have seen

before; and 3) the assessment of the quality of information items cannot be achieved, like distinguishing a well written article and a badly written article.

Collaborative filters are therefore introduced to help people make choices of the Web information blocks based on the opinions of other people. The issue was raised when information resources contain a large number of information items with different subjects and of various interests making it difficult for people to obtain the items that they really want and for information providers to protect the copyrights of the items they want. There are several papers devoted to this subject [Maltz95, Shardanand95, Resnick94].

Some advanced filtering approaches [Shardanand95, Maltz95] have been suggested to overcome the limitations of the basic filtering approach. The basic ideas of the approaches are 1) to maintain a user profile or record of the user's interests in specific items, 2) to compare the profile with the profiles of other users and to weigh each profile for its similarity with the user's profile, and 3) to define a set of the most similar profiles and use information contained in them to recommend items to the users.

A second approach to filtering information into a group of people of similar interest (e.g. netnews group) is to correlate the ratings of articles, assigned by users to articles, in order to determine which users' ratings are most similar to each other, and then to predict how well users will like new articles, based on ratings from similar users. The approach is based on the assumption that people who agreed in their subjective evaluation of past articles are likely to agree again in the future.

As we can see from the above, a basic idea in information filtering lies in a well defined formulation and calculation of similarity in the Web information items. Here the most important factor is what attributes are selected for similarity computation and how to select them. We will discuss this important issue in the next sections.

2.3 Metadata

PICS is an infrastructure for associating labels (metadata) with Internet content. It is a platform toward interpretation of both semantic and syntax of information content being documented. How to manage the content labels and labelling appears a difficult task because not only are they embedded in contents, not easy to pick out, but also no unified explanatory word set is available for the data.

There is no metadata modelling support available for describing content labels and labelling. Most of PICS work up to date is no more than a contemporary attempt at finding a modelling approach to metadata problems. Some features can be drawn for describing metadata. They are, among others, heavily human interactiveness, required semantic interpretation, multiple representations (multimedia), format and content changeability and pragmatic-proneness.

Keys to metadata formulation lie in a sufficient selection method for content labelling, where the contextual semantics can be acquired by figuring or pointing the labelled words. Metadata, in a sense, is a connotation of words labelled to basically interpret what the related contexts mean. A well defined metadata formulation may be, with help of sufficient vocabulary, providing an appropriate means for filtering information to be searched.

Substantial information to be searched through labelled contents may give rise many semantic conflicts, i.e., duplicated blocks of information or contradictory blocks of information found. Detection of such conflicts depends as well on a well defined metadata formulation. In the next section, we will give detailed discussion on this aspect.

2.4 Related Activities

2.4.1 Copyright protection

Unlike conventional paper-based material, the copy and distribution of digital information involves very little effort and almost no cost. Because of the enormous amount of such information that can be freely reproduced and distributed on the Internet, some recent activities have been concentrated on mechanisms to protect the intellectual property of authors and content providers.

Although the primary goal of PICS is one of content filtering, it has been suggested that the labels that are part of the PICS technology could contain some information regarding copyright ownership, distribution rights, and even requested payments. In this sense, software components implementing PICS could check for such labels and demand payment before distributing the labelled items.

2.4.2 Guided search

One of the main assumptions for establishing collaborative filters is to use previous readers' marks on a certain information item. The marks are used by statistics to form a basic evaluation toward a subject. New comers, when presenting their interests, will be classified according to this evaluation formulation. For instance, in a news group or a bulletin board, there are a number of subjects or topics used for classification and identification of massive information. A rating service is employed for identifying the closeness among the subjects. When a new subject or topic is introduced, it can be either added as a sub-subject of a certain subject or as a new subject. How to manage this addition of the new subject or topic is the task of a rating mechanism provided by a rating system.

2.4.3 Vocabulary guided and control

Another way of content filtering is to provide vocabulary with the Web documents. That is, when a document is sent to the Internet for publication, a set of words (called vocabulary) should be simultaneously provided for identifying the document. These words associated with the document are in general sufficient to tell the main content of the document. There are three sources, from which vocabulary can be obtained.

The first way of obtaining content vocabulary is from the information providers. When an information item is sent to the Internet, a set of words should be included in the information item for readers to identify what sort of information is provided. Then by using the set of words (vocabulary) readers can decide whether the information item is interesting to them or if it should be blocked from other readers. The problem is that the information providers with or without purposes ignore the attachment of vocabulary to the information items.

The second one is to define a set of standard vocabulary which can be applied by all net readers. Net supervisors (like news group organisers and parents) can check the Web information (in general, the topics and key words of the information documents) which are being accessed by readers against the standard vocabulary and decide whether the documents are allowed or needed for the readers. The problem is that the standards cannot meet different tastes of readers as well as supervisors. That is, the vocabulary can be too tight a restriction on one group of readers and too loose on the other.

The third way is to set up vocabulary by supervisors themselves (like parents and organisation heads). The problem here can be that they have to learn what information is likely to be accessed through the Internet. The degree of appropriateness of the Web materials to their children, for example, may vary from parent to parent. In addition, it may be easier for parents to adopt a filtering program to block harmful materials from children under 6, than from those between 16 and 10, who may have access to the harmful materials by going around the block facility.

3 PICS Issues

3.1 PICS platform design

It is described in [Resnick96] that PICS establishes Internet conventions for label formats and distribution methods while dictating neither a labelling vocabulary nor who should pay attention to which labels. The most important components of a PICS document are:

1) a syntax for describing a rating service – so that computer programs can present the service and its labels to users,
2) a syntax for labels – so that computer programs can process the labels,
3) an embedding (list) of labels in the RFC-822 transmission format and the HTML format – so that the labels can be included in Web page contents,
4) an extension of the http protocol – so that clients can request that labels be transmitted with a document, and
5) a query-syntax for an on-line database of labels.

These components of a PICS document should be considered when designing a rating system which supports any access control or guided searching.

3.2 Guided Search and Access Blocking

3.2.1 Hierarchical management of Web contents

Governmental initiatives in the US and other countries have been issued to provide searchable information about the content of governmental agencies' information archives (GILS), where a hierarchical management of the Web information is provided.

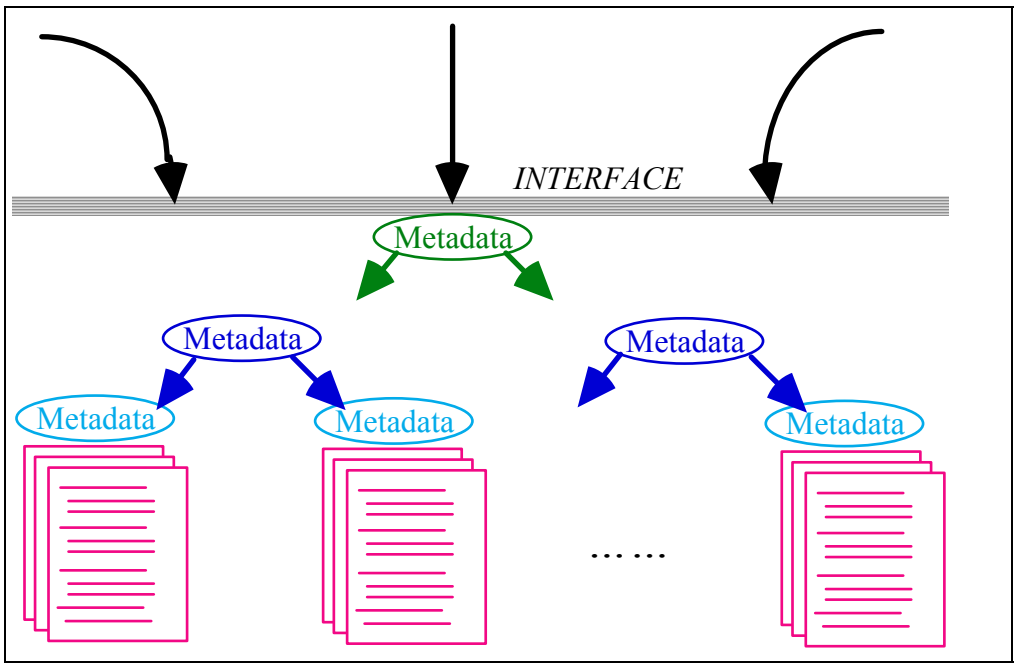

Fig 2. A hierarchical management of metadata.

In this structure, Web information documents are organised in terms of their metadata structures, which will be described in the following section. One metadata instance contains labels from a particular document and URL (if needed) where the document is located. The instance in general gives semantics of this particular document. Related metadata instances will again be grouped into a higher level of metadata schema, which describes properties of the grouped Web information documents.

The structure is accessed by the Web users through a special interface or searching engine, which will guide the users to find what information they look for. A key factor is that the structure should be supported by a vocabulary mechanism so that new information documents input can be sorted and inserted into a suitable place in the hierarchical structure.

3.2.2 Blocking accesses

Computer software can implement access controls that take into account all these factors. The basic idea, illustrated in Fig 3, is to interpose selection software between the recipient and the on-line documents. The software checks labels to determine whether to permit access to particular materials. It may permit access for some users but not for others, or at some times but not others.

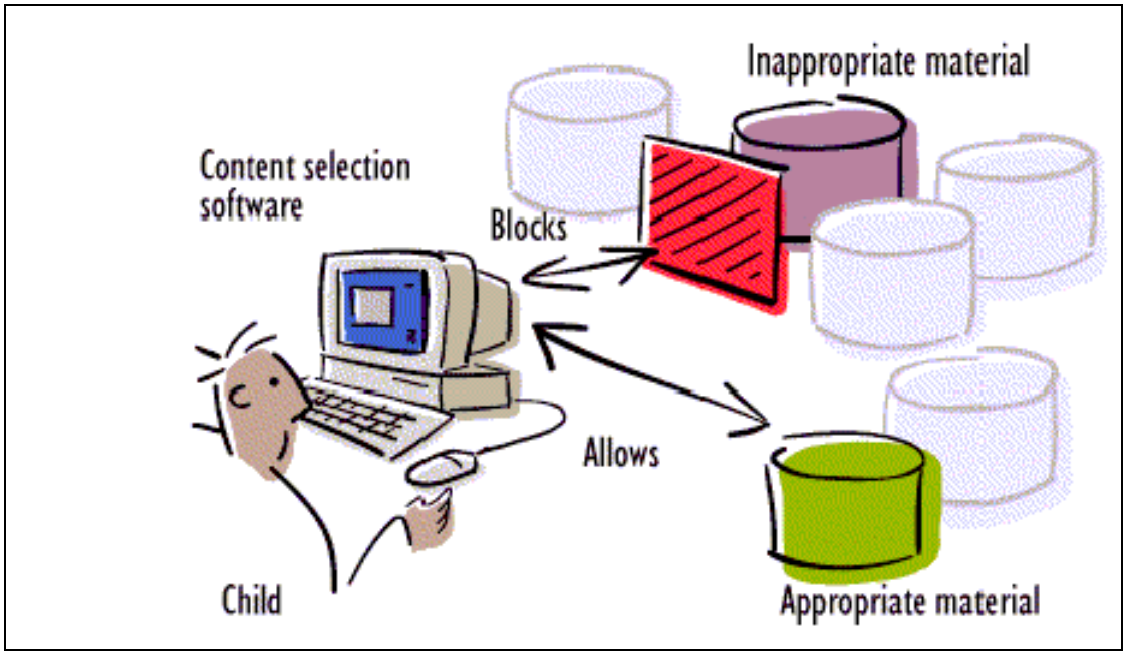

Fig 3. Selection software automatically blocks access to some documents.

PICS provides a common format for labels, so that any PICS-compliant selection software can process any PICS-compliant label. A single site or document may have many labels, provided by different organisations. Consumers choose their selection software and their label sources (called rating services) independently, as illustrated in Fig 4. This separation allows both markets to flourish: companies that prefer to remain value-neutral can offer selection software without providing any labels; values-oriented organisations, without writing software, can create rating services that provide labels.

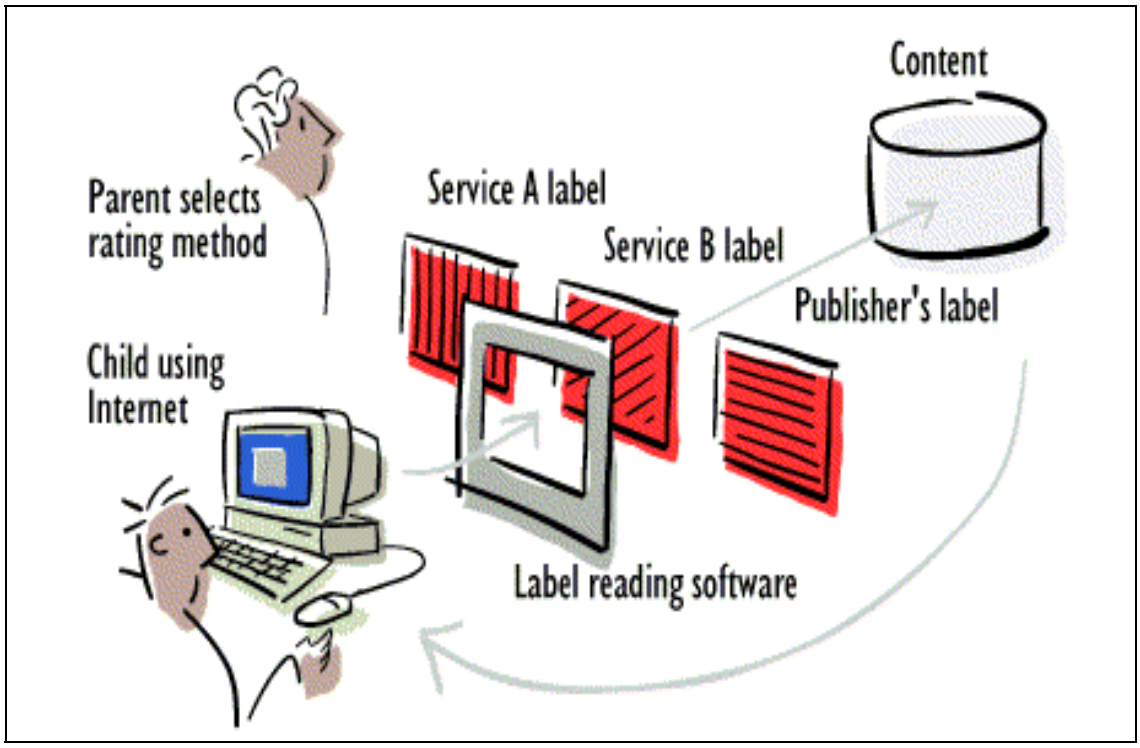

Fig 4. Selection software blocks based on labels provided by publishers and third-party labelling services, and on selection criteria set by the parent.

3.3 Metadata

3.3.1 Metadata transformation

The concept "metadata" can be viewed as a semantic focus of a selected information item which is being labelled or accessed through URL. However, the semantics of the content item labelled are determined by the information provider according to how he or she understands the content. His or her selection is usually a subjective consideration on the content based on the knowledge frame he or she perceives. Therefore, there is no "standard" schema of metadata which applies to general content selection and, moreover, such schema of metadata would make communication quite difficult.

Due to the transformation problem between different metadata schemas used in retrieval systems, a logic form and syntax of metadata, called Multi-Schema Metadata Format (MMF) is defined to accomplish the inter-schema transformation of metadata [Sakata97].

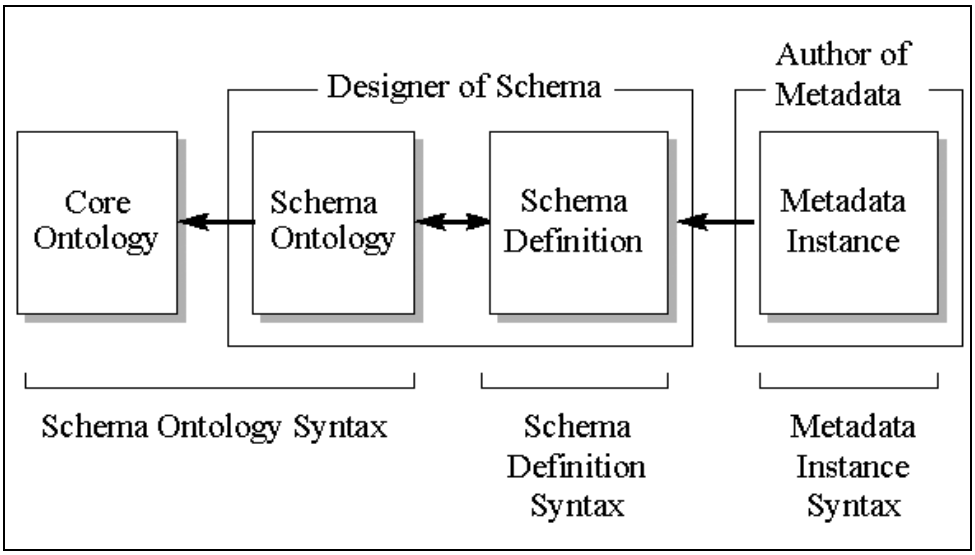

Fig 5. Multi-Schema Metadata Framework.

The Multi-Schema Metadata Format (MMF) consists of four components: metadata instance, schema definition, schema ontology, and core ontology. A metadata instance comes from a HTML description of information (objects) by a Web author. The description usually contains attributes of objects to be described and their values. The set of attributes and their values in HTML is then extracted and mapped into an attribute schema (like a data schema), where relationships between attributes are defined (and form relations or entities in a relational model). Then attribute schemas (entities, if we view them in an ER model) are related to each other to generate a basic description of schemata (or a model). Finally, as the number of schemata increases, it is necessary to have a model description which can standardise schema design.

The merit of MMF provides a support that the Web information items can be formulated and analysed by being mapped into a well-defined and formalised schema description, and hence transformed between different formats in "home languages (defined by the Web authors).

3.3.2 Metadata formats

Metadata is data which describes attributes of a resource. It basically supports a number of functions, such as location, discovery, documentation, evaluation, selection, and so on. It is believed that in an almost infinite resource space, effective management of networked information will increasingly rely on effective management of metadata.

Of course, effective management of metadata mainly depends on how well the metadata is structured. It is suggested that three classes can be defined based on the structuralness of metadata (i.e., how well authors structure metadata). The first class includes relatively unstructured data, merely automatically extracted from resources and indexed for searching. The data has little explicit semantics and does not support field searching.

The second class includes data containing enough description that users are allowed to assess the potential use or interest of a resource without having to retrieve it or connect to it. The data is structured and supports field searching. The data description usually gives discrete objects instead of capturing inter-object relationships.

In the third class, full descriptive formats are used to not only locate and discover objects, but to play a role of documenting and collecting objects as well. The data is expressive to capture a variety of relationships between objects at different levels.

4 Some Applications

In this chapter, we suggest two scenarios for PICS applications with focus on the Web access control. A good story about parenting described by P. Resnick is directly used in the section 4.1.2 for describing how a parent sets access control on her Web browser for supervising her children to navigate the Internet information. As an application example, we give a PICS program to specify how a publisher applies labels in a rating system to support access control in its software publication organisation.

4.1 Scenarios

4.1.1 Accessing to a book

The following is a short example to show that "Access Right" to a book can be realised in terms of PICS format. Detailed implementation can refer to the example program in the next section.

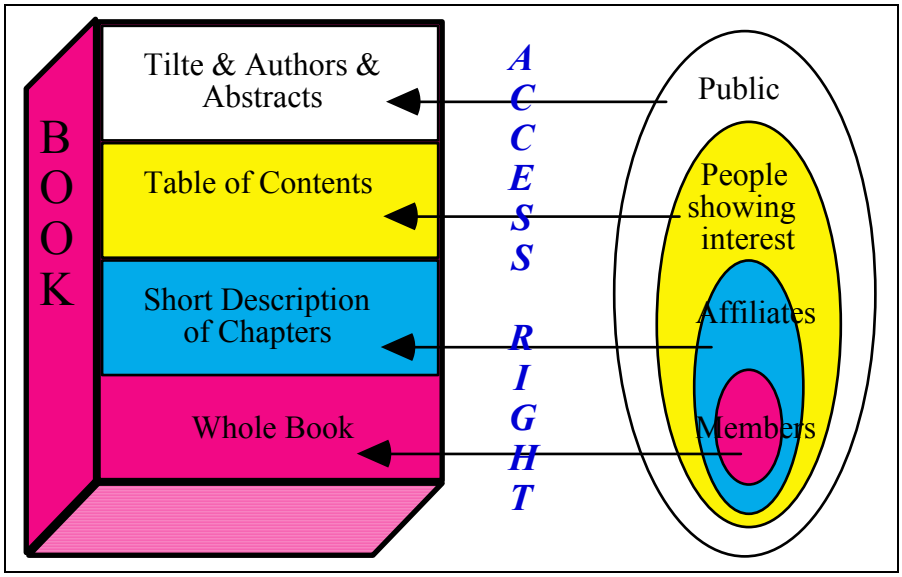

Fig 6. Access right of people to a book.

The above figure, Fig 6, shows a possible way of using PICS rating service to provide access control to a project book, of which, different parts are available to different reader groups. That is, different groups of people have different access rights to the book. For instance, the members of the project have the right to access to the whole book, while public people only can have a look at the title, the authors, and the abstracts of the book.

4.1.2 A parenting story

A parent (for the sake of concreteness, we will assume the mother) would like her son, age 10, and daughter, age 14, to explore the wealth of information that is available through the Internet. She is concerned, however, that the children may retrieve text or images or participate in interactive chat sessions that they are not mature enough to

handle.

1. The parent installs labelling-compliant browser software on the home computer.

2. Using the computer, the parent reads about several labelling services. Some services charge a fee for their use, while others are free. Each service describes its labelling system and illustrates it with examples. For example, some services rate only on the presence of nudity and have very concrete criteria, while others take into account multiple factors. Some services rely on computer analysis of text and images to assign labels, while other services employ people to assign the labels.

3. The parent finds that a service called GoodCleanFun is most compatible with her values. GoodCleanFun employs people to label items by the minimum age at which children should, according to its values, be exposed to the items.

4. The parent configures the browser so that it checks with the labelling service before downloading text or pictures, or connecting to chat channels. She configures it so that her 10 year old son can access only information that GoodCleanFun marks as appropriate to his age group. She believes that her 14 year old daughter is mature for her age and can safely be exposed to materials labelled for 16 year olds; the parent configures the browser software accordingly.

5. GoodCleanFun attempts to label everything available on the Internet, but inevitably misses some things. The parent configures the browser to block the younger child's access to items that GoodCleanFun has not labelled. For the 14 year old, she configures the browser to consult the publisher's own labels. The parent is not thrilled with the rating criteria used by most publishers to label their own content, and she knows that some publishers deliberately mislabel things. Still, for those items that GoodCleanFun has not labelled, she decides that the publisher's labels are an acceptable backup.

6. Either child may use the computer without the parent present, but the browser software always checks GoodCleanFun's labels and enforces the rules that the parent has set. The browser may either retrieve the labels from GoodCleanFun over the Internet, or read them from a CD-ROM.

7. The browser may indicate to the child which links are available, perhaps with a colour coding scheme. If the child attempts to retrieve something that is blocked, the browser offers an explanation of why the item is blocked. At the child's urging, a parent can enter a password to temporarily override the block.

4.2 Application example

The following is an example of application for labelling service. Here the rating service is provided by the organisation of US Software Publisher Group (SPA) and the rating system is described in HTML: "http://www.spa.org/publishersv01.html". The label settings suggest four kinds of access codes, valued from 0 to 3, for "Not Member", "Individual", "Corporate Member", and "Applet Publisher", respectively. More document categories (in HTML format) can be defined for different document type description.

```
((PICS-version 1.0)
        (rating-system "http://www.spa.org/publishersv01.html")
        (rating-service "http://www.spa.org/")
        (name "Software Publishers Association Members")
        (description "The SPA is a group of US Software Publishers. Its members
        adhere to the highest professional standards in producing consumer software")
        (category
                (transmit-as "m")
                (name "Status as a Software Publisher")
                (label
                        (name "NOT Member")
                        (description "Firm is NOT a member of SPA")
                        (value 0))
                (label
                        (name "Individual")
                        (description "Though the SPA does not have individual members,
                        it has a signed pledge from this individual on file.")
                        (value 1))
                (label
                        (name "Corporate Member")
                        (description "SPA Member in good standing as of label_time")
                        (value 2))
                (label
                        (name "Applet Publisher")
                        (description "SPA Member which has been certified as an Applet
                        publisher and signed a pledge of writing secure software")
                        (value 3))
        )
)
```

5 Future work

Many issues in using the Web information remain insufficiently explored. One reason can be that people prefer practical applications of the Web to finding out what theoretical support. However, as applications and researches in the Internet proceed, people start considering to solve some urgent problems like structuring the Web information. Among these issues, we believe that metadata, semantics of the Web information, and integration support could be interesting topics and likely to be addressed.

In the following, we give a short description of the three points to illustrate what we attempt to analyse in the next step.

Metadata Framework

Initial Web information is considered to be not well structured or even unstructured. The Web data normally in textual lines are not organised in objects and relations and therefore contain no explicit semantics. This situation of the Web information generation makes it difficult to parse the textual lines and to analyse the connotation of underlying contents. However, if a mature modelling support can be applied for analysis of the Web contents, the difficulty of direct parsing of the Web contents may become easier to overcome.

Here a key point is to extract the Web data, to define or choose a suitable metadata framework, and to transform the Web data, along with their frameworks into a modelling support method which has the capacity to analyse the Web data and capture the semantics of the Web data. To start with, one such modelling method can, for example, be Relational Model, by which Web information items can be organised into attributes and relations.

Semantics knowledge representation

In control of access to the Web information, the assessment of what information to obtain and what information to block is fully a subjective attribute. A user's profile may provide a sort of semantic framework for selection of the Web information items, but unstructured Web data containing no explicit semantics may appear too difficult to be semantically captured. A key point is to define a formal representation, i.e., a semantic framework, for the Web information. This representation will support semantic relationships between labels (metadata) and similarity comparison between labels.

Integration issues – Similarity and Inappropriateness

As we have briefly discussed, the activities to group the Web information items for Newsgroup or Bulletin Board and to detect what information is inappropriate for the Web readers require understanding and analysis of the Web information items, as well as computation of semantic similarity and dissimilarity of the Web information items. Once the Web information is extracted and represented in a semantic framework (or a semantic model), to analyse and compare the semantics of the Web information items (can be viewed as objects) will have to be performed. For a newsgroup, the Web information items with high similarity (hard to judge) will be grouped together while for access blocking, an information item, while semantically similar to the blocking settings (which can be viewed as a semantic framework) by parents, will be blocked.

The key issues are to find semantically comparable items from the Web information items and to form semantic schemas for representing the information items or objects.

References

Resnick, P. and J. Miller, PICS: Internet Access Controls without Censorship, CACM, Vol.39, No.10, 1996

Maltz, D. and K. Ehrlich, Pointing The Way: Active Collaborative Filtering, CHI'95, 1995

Shardanand, U. and P. Maes, Social Information Filtering: Algorithms for Automating "Word of Mouth", CHI'95, 1995

Weinberg, J., Rating the Net, 1995

INFO2000, Illegal and harmful content on the Internet, 1996

Resnick, P., etc. GroupLens: An Open Architecture for Collaborative Filtering of Netnews, Proc. of ACM conf. on Computer Supported Cooperative Work, 1994

Sakata, T. etc., Metadata Mediation: Representation and Protocol, Proc. of WWW6, Santa Clara, U.S.A., 1997

Isberg, G., The ethics, the responsibility and the supporting tools (in Swedish), Om Skoldatanätet, Project report, 1997

Kulturdepartementet, Grönbok om skyddet av minderåriga och den mänskliga värdigheten inom de audiovisuella tjänsterna och informationstjänsterna (in Swedish), March, 6 1997.

Miller, Jim, etc., Rating Services and Rating Systems (and Their Machine Readable Descriptions), Version 1.1, W3C Recommendation 31-October-96.