

## DESIRE Information Gateways Handbook

[HOME](#)
[TABLE OF CONTENTS](#)
[AUTHORS](#)
[Search](#) | [Help](#)

### Information Gateways Handbook (Print Version)

#### Section 1 : Strategic Issues (Print Version)

##### Target audience

---

Section 1 of this handbook is aimed at the people responsible for strategic management - funders and project managers who will initiate the set up of a gateway and who will steer its direction over time.

It aims to give an overview of the key issues involved in gateway projects, giving a rationale for these projects. It covers the important decisions that need to be made when setting up a new gateway (for example, staff effort, skills and costs) but also deals with logistics for managing an existing gateway.

Each section offers some background, practical tips and hints, key references, a glossary, case studies and examples. Watch out for the  **CROSS REFERENCE** that will take you to related sections elsewhere in the handbook

##### Contents

---

Section 1 : Strategic Issues

1. [Information Gateways overview](#)
2. [Preliminary planning](#)
3. [Staff and skills required overview](#)
4. [System requirements overview](#)
5. [Maintenance requirements:cost implicaitons](#)

Section 2 : [Information Issues](#)

Section 3 : [Technical Issues](#)

#### 1.1. Information gateways overview

##### In this chapter...

---

- what is an information gateway
- the rationale for developing information gateways
- examples of leading information gateways

##### Introduction

---

Information gateways are now a well established feature on the Internet. There are a number of different models for setting up and running gateways. The technology behind gateways can also vary considerable. But quality information gateways all have key similarities that make them invaluable resources to their respective user communities.

##### What is an information gateway?

---

Information gateways are quality controlled information services that have the following characteristics:

1. an online service that provides links to numerous other sites or documents on the Internet
2. selection of resources in an intellectual process according to published quality and scope criteria (this excludes e.g. selection according to automatically measured popularity)
3. intellectually produced content descriptions, in the spectrum between short annotation and review (this excludes automatically extracted so-called summaries). A good but not necessary criterion is the existence of intellectually assigned keywords or controlled terms.
4. intellectually constructed browsing structure/classification (this excludes completely unstructured lists of links)

5. at least partly, manually generated (bibliographic) metadata for the individual resources

After T. Koch: <http://www.ub2.lu.se/tk/SBIG-definition.txt>

### **The rationale behind information gateways**

Many academic libraries and institutions are currently looking for ways to help their users discover high quality information on the Internet in a quick and effective way. The **DESIRE** project and others (e.g. [IMesh](#)) suggest that the development of information gateways can provide a solution.

Researchers and academics do not always have the time, inclination or skills to surf the Internet for resources that could support their work. As Internet publishing and communication become more commonplace this could disadvantage some researchers as they will miss valuable information and communication resources.

In the traditional information environment human intermediaries, such as publishers and librarians, filter and process information so that users can search catalogues and indexes of organised knowledge as opposed to raw data and disparate information. Subject gateways work on the same principle - they employ subject experts and information professionals to select, classify and catalogue Internet resources to aid search and retrieval for their users. Users are offered access to a database of Internet resource descriptions which they can search by keyword or browse by subject area. They can do this in the knowledge that they are looking at a *quality controlled* collection of resources. A description of each resource is provided to help users assess its origin, content and nature, enabling them to decide if it is worth investigating further.

### **Examples of leading information gateways**

The following information gateways are used elsewhere in the handbook as examples of good practise and/or having interesting development information to contribute to the wider gateway's community. A full listing of information gateways can be obtained from:

- <http://www.hw.ac.uk/libWWW/irn/pinakes/pinakes.html>

#### **E X A M P L E**

##### **Leading information gateways**

##### **Biz/ed - Business and Economics Education on the Internet**

Biz/ed is a unique business and economics service for students, teachers and lecturers. The gateway contains a ROADS based Internet catalogue with over 1400 Internet resources selected and described by subject experts.

- <http://www.bized.ac.uk/>

##### **DutchESS - Dutch Electronic Subject Service**

Is an Internet Subject Service which indexes Internet resources, selected on quality and relevance for the academic community: students and academic researchers. The resources are classified according to the Nederlandse Basisclassificatie (Dutch Basic Classification).

- <http://www.konbib.nl/dutchess/>

##### **EEVL - The Edinburgh Engineering Virtual Library**

The EEVL Service a gateway for the higher education and research community to access high quality information resources in Engineering. The EEVL gateway offers broad or focused searching capabilities, and search results provide the choice of linking to full descriptive resource records or to the resources themselves. The catalogue has descriptions and links to thousands of quality Internet resources.

- <http://www.eevl.ac.uk/>

##### **The Finnish Virtual Library Project**

The Finnish Virtual Library project, launched in 1995 and funded directly by the Finnish Ministry of Education, aims to form a foundation for a Finnish field-specific subject index of subject gateways. A collection of libraries have produced individual virtual libraries in 40 subject areas; these are now being converted into a gateway format, and offered as bilingual services in Finnish and English. The Kuopio University Virtual Library has mounted its Virtual Library as a ROADS-based gateway, covering the subject areas of Clinical Nutrition, Neurosciences and Pharmacy.

- <http://www.uku.fi/kirjasto/virtuaalikirjasto/>

#### **NMM Port**

Port is the UK National Maritime Museum's online catalogue of high quality maritime related Internet resources. Every resource has been selected and described by a librarian or subject specialist. Services and materials developed by the Museum's Centre for Maritime Research are also available on the site.

- <http://www.port.nmm.ac.uk/>

#### **OMNI - Organising Medical Networked Information**

OMNI, Organising Medical Networked Information, covers the areas of medicine, biomedicine, allied health, health management and related topics. The service also provides training materials and workshops. Browsing can be done via either alphabetical topics, classified topics, or via MeSH headings. In addition, OMNI provides a range of biomedical value-added services, including a MEDLINE review section, mirrors of key NHS IT strategy documents, and the UK CME database.

- <http://www.omni.ac.uk/>

#### **SOSIG - The Social Science Information Gateway**

SOSIG can help you locate high quality sites on the Internet, which are relevant to social science education and research. The Internet Catalogue offers access to thousands of high quality Internet resources, each selected and described by academic librarians and subject specialists. The SOSIG service receives funding from the ESRC, JISC and the European Union.

- <http://www.sosig.ac.uk/>

---

### **Glossary**

**Desire** - Development of a European Service for Information on Research and Education, EU funded research project

**ESRC** - Economic and Social Research Council. The ESRC is the UK's largest independent funding agency for research and postgraduate training into social and economic issues.

**IMesh** - International Collaboration on Internet Subject Gateways

**JISC** - Joint Information Systems Committee. UK Higher Education organisation, with the aim to stimulate and enable the cost effective exploitation of information systems and to provide a high quality national network infrastructure for the UK higher education and research councils communities

**ROADS** - Resource Organisation And Discovery in Subject-based Services

---

### **References**

Biz/ed - Business and Economics Education on the Internet, <http://www.bized.ac.uk/>

Desire - Development of a European Service for Information on Research and Education, <http://www.desire.org/>

DutchESS - Dutch Electronic Subject Service, <http://www.konbib.nl/dutchess/>

EEVL - The Edinburgh Engineering Virtual Library, <http://www.eevl.ac.uk/>

The Finnish Virtual Library Project, <http://www.uku.fi/kirjasto/virtuaalikirjasto/>

IMesh, <http://www.desire.org/html/subjectgateways/community/imesh/>

NMM Port, <http://www.port.nmm.ac.uk/>

OMNI - Organising Medical Networked Information, <http://www.omni.ac.uk/>

PINAKES - A Subject Launchpad, <http://www.hw.ac.uk/libWWW/irn/pinakes/pinakes.html>

SOSIG - The Social Science Information Gateway, <http://www.sosig.ac.uk/>

## Credits

---

Chapter author: [Martin Belcher](#)

Contributors: Phil Cross

## 1.2. Preliminary planning

### In this chapter...

---

- setting a gateway's objectives
- examples gateway objectives
- scheduling achievable timescales
- phasing of the project

## Introduction

---

Information gateway projects range in size and complexity from small scale projects, that an enthusiast embarks upon in their own time, to the development of full blown services at a national level, that a team of many specialists works on full time. This handbook is primarily concerned with the development of larger scale gateways. This chapter deals with the planning of a medium to large scale gateway and not a "one-man" band approach. Saying that, many of the issues that are applicable to a large scale gateway are equally applicable to a gateway set up by a single person. However, the system of a well defined plan, aims and objectives, and a carefully thought out timetable should help contribute to any gateway project, regardless of its size.

## Background

---

As with any serious project, a well thought out plan is essential for long term success of an information gateway project. The best way to plan projects efficiently is with the aid of a formal project plan document. An important section of the project plan is a clearly defined set of aims and objectives. Simply stating what a project's aims and objectives are is not enough. The objectives must be accompanied by a clear set of deliverables, against which the overall success of meeting the aims and objectives can be measured. The deliverables need to be contextualised with a clear and simple timetable to help deliver the project within a sensible time frame.

## Setting a gateway's objectives

The fact that you are seriously considering setting up a gateway must mean that you have some aims and objectives. This might be to establish a service for a specific national user community, or perhaps it is to set up a gateway for your University Library? Each different gateway will have a different set of aims and objectives. If you are receiving funding from a third party then it is highly likely that there are some contractual aims and objectives that have to be met.

In general aims and objectives are wide ranging and rather broad statements that require further clarification. A measurable set of scheduled deliverables can help focus the general aims and objectives. Deliverables are an important part of a project plan and are often required as a condition of funding (it allows the funding and supporting organisations to check and evaluate that their funding is being used to achieve the project's set aims and objectives.)

### EXAMPLE

#### Early SOSIG project aims and objectives

An early SOSIG project plan (published February 1996) contained the following text:

SOSIG's overall aims fall into three broad categories:

- To improve delivery of information and quality of service by working with and helping to pilot the latest developments in networked resource tools technology
- To improve accessibility and usability of resources via a programme of training and awareness
- To encourage availability of new, quality networked resources of relevance to social scientists

Social Science Information Gateway - Project Plan  
(Lesly Huxley and Nicky Ferguson: 1996)

#### Early SOSIG deliverables

Also contained in the same document, were a set of key deliverables that helped to put the broad aims and objectives into easily measurable deliverables. A sample of early SOSIG deliverables include:

- A demonstrator service providing a testbed for the latest developments in networked information retrieval technology in collaboration with other services
- Subject-specific training documentation (in paper and online form)
- Subject-specific training workshops
- Subject-based user guides to selected quality networked resources
- Promotional materials to raise awareness of the service

Social Science Information Gateway - Project Plan  
(Lesly Huxley and Nicky Ferguson: 1996)



### REMEMBER

#### Deliverables should be SMART:

- **S**pecific
- **M**easurable
- **A**chievable
- **R**elevant
- **T**ime-based

Making your deliverables SMART can help everyone involved in the project, both those involved in the implementation and those involved in the funding of the project.

## Scheduling achievable timescales

---

Once a detailed set of deliverables has been drawn up, the next stage is to develop a timetable for their delivery. There are a few issues to consider when committing to a timetable, the most important issue being that once you have an agreed timetable then you are bound by it. There may be some flexibility in the schedule, but generally deadlines should be kept to, in order to avoid projects running into timetabling difficulties. Therefore developing a realistic and achievable timetable is important.

There is little point in having lots of important sounding deliverables and a very detailed timetable if the schedule is impossible to meet. It is a guaranteed way to increase the chances of the project and hence the gateway, failing. Set realistic and achievable deliverables and deadlines. Do not agree to do something unless there is sufficient time and resources available to deliver.

## Phasing of the project

---

Many of the tasks associated with setting up an information gateway are closely related to each other. There is an overlap with some tasks whilst some can only be started once others have been completed. The key tasks and phases of an information gateway project might include:

### Phase 1: Pre-project

- Outline planning of project
- Securing funding for project
- Producing outline project timetable and plan

### Phase 2: Project planning and set-up

- Drawing up detailed timetable and plan
- Hiring staff and developing skills
- Developing policy documents (scope and selection criteria)
- Technical planning

### Phase 3: Technical implementation

- Technical set up and system testing
- Training of non-technical staff in system usage

### Phase 4: Catalogue development

- Cataloguing of resources and catalogue development
- Service launch

### Phase 5: Day to day running

- Ongoing catalogue development
- Collection management

Generally the phases above are all sequential and related i.e. phase 3 can't really be started until phase 2 has been completed, etc. The actual launch date of the gateway should often be delayed until there are a certain number of resources in the catalogue. Many gateways have waited until 100-200 resources are available before launching. Although the exact number will be largely dependent on the staff effort available to develop the catalogue and the overall objectives of the gateway.

## References

---

SOSIG, <http://www.sosig.ac.uk/>

## Credits

---

Chapter author: [Martin Belcher](#)

## 1.3. Staff and skills required overview

### In this chapter...

---

- setting up a gateway
- running a gateway
- skills and people checklist

### Introduction

---

Information gateway projects have several distinct phases; planning and scoping, technical and information setup, administration and maintenance. Each phase requires different skills and perhaps different staff. In the ideal world a gateway project would be able to call on a large pool of staff, this may be the case in some instances, more often a few key staff will perform the majority of the tasks, with external people being brought in from time to time.

### Setting up a gateway

---

Depending on the exact technology used, there is going to be relatively large up front cost in terms of time and unique skills, in the setting up of a gateway. The information management issues will require research and documentation. It is likely that the people involved with this side of the setting up, will continue to play a part in the project, most usually in the building of the resources database and the day to day running of the project. There will also be a large up front cost in terms of the technical implementation of the infrastructure software that the gateway will operate on. How large this cost will be depends on whether or not an existing set of gateway technology is being used (e.g. [ROADS](#)) or a new system is being developed. Either option will require people with the appropriate technical skills.

If the gateway technology is being developed from scratch or using an existing system with significant modification, then significant amounts of technical research and development will be required. Staff with the appropriate technical skills will be essential. Additionally there may be a need for an interface designer, to develop the user front end to the system. These skills will only really be required for a set period and set of tasks. As such they are the ideal skills to bring in from external sources.

A project manager or supervisor will also be invaluable, to help in the development of the project to time, budget and its original aims and objectives. The project manager should be able to operate on both the subject specialist level and technical level. This doesn't mean that you need a programming librarian, but someone who can understand both areas and manage their different strengths and weaknesses.

### Running a gateway

---

The key staff needed for the running of a gateway are subject specialists who will be involved in the expansion and development of the resources catalogue. The exact number of these will depend on the scope of the gateway. If the gateway aims to catalogue all resources in a given field within a short period, then a larger number of cataloguers will be required. The more subject specialist and resource cataloguers there are, then the faster the number of resources in the gateway can grow.

Various models of developing the catalogue of resources and distributed staffing are discussed elsewhere (resource discovery strategies, working with information providers and distributed cataloguing and collaborative working), each model can have a significant effect on the number and type of core staff that a gateway requires for expanding the catalogue of resources.

#### CROSS REFERENCE

[Resource discovery](#), [Working with information providers](#), [Distributed cataloguing](#), [Co-operation between gateways](#)

Depending on the technology used to set up and run a gateway, the need for continued technical support and development can vary considerably. Under some circumstances the need for technical support staff effort can be kept very low. However, it is essential for the long term survival of the

gateway that a reasonable amount of staff effort is kept aside for technical support and development. Even the most robust technologies can run into problems. Simple problems can cripple a gateway if the technical staff are not there to fix them.

### Skills and people checklist

Under ideal circumstances an information gateway will be able to draw on the skills of staff with the following roles and/or job titles. Reality may mean that a few staff cover all these roles:

Title	Description	Skill Set
Project manager	someone to over see the whole project and ensure the smooth day to day running	organisational skills, good written and oral communication, person management, subject and technical knowledge and understanding, excellent information management skills
Subject specialist	person or persons to develop the intellectual scope of the gateway and the expansion of the gateway catalogue or resources	excellent subject knowledge, understanding of information management issues, ideally extensive Web experience and some understanding of technological principles behind gateway
Information cataloguers	person or persons directly involved in the entry of resources into the catalogue (often the same as the subject specialist)	subject knowledge, confident Web user, some understanding of technological principles behind gateway
Technical implementation officers	person or persons involved in the development and implementation of the technical side of the gateway	excellent technical understanding of the networked environment, good programming and scripting skills and good working knowledge of proposed gateway technology. If developing new gateway technologies then very high network related technical skills are essential. Ideally have some appreciation of information management issues
Technical support officers	person responsible for the day to day technical integrity of the gateway system	as technical implementation officers but can be slightly less experienced if correct tools are put in place in the system development
Web server administrator	person responsible for the running and administration of the gateway web server	as above plus excellent Web server administration skills
User interface designer	person or persons responsible for the design and implementation of the gateway user interface	good understanding of Web site design and well versed in usability and accessibility issues
Finances officer	person responsible for the financial side of the project	good understanding and experience of potentially large scale project financial management, may or may not be project manager
Publicity and promotions officer	person or persons responsible for the development and deployment of publicity and promotional materials/activities	experience in publicity and promotions, good subject knowledge and user community understanding

### The ideal versus the real world

Ideally we would all like to be able to draw on the specialist skills of all those people outlined above. The real world dictates that more often than not, we will be required to draw the skills from a smaller group of multi-skilled people. This means a very broad skill set is required from a small number of staff. It can also mean the development of an excellent, tight-knit, well focused team.

When skills are lacking within the core team, it can often be very effective to bring in experts from outside. These experts could be drawn from within the same organisation (e.g. other sections of the same university) or they could be commercial consultants. People involved in the technical



implementation, user interface design and publicity and promotion are often brought in under such circumstances.

## Glossary

---

**ROADS** - Resource Organisation And Discovery in Subject-based Services

## Credits

---

Chapter author: [Martin Belcher](#)

## 1.4. System requirements overview

### In this chapter...

---

- reliability - making sure your gateway is always available
- responsiveness - how will your gateway perform?
- efficiency - making the best of available resources
- scalability - coping with more users, more data and more services

## Introduction

---

Subject gateway services need to be provided in such a way that they are:

- reliable
- responsive
- efficient
- scalable

A reliable service is one that is available all (well, almost all) of the time, is secure and does not lose all your data in the event of disk failure or security breaches. A responsive service is one that can be browsed, searched and maintained in a way that does not subject the end-user and cataloguer to undue delays. An efficient service makes the best use of the available hardware and network resources. A scalable service is one that can cope with demands placed on it by growing numbers of end-users, increasing database size and new service scenarios.

## Background

---

Subject gateways operate in a Web environment. This means that they must be available all the time. End-users expect reasonable response times while they browse the gateway and fast and predictable performance when they search the database. Subject gateway cataloguers expect reasonable response times as they add resource descriptions to the database. Subject gateway managers want to be able to deliver all this at a reasonable cost - both in terms of the initial cost of establishing the gateway and in terms of ongoing hardware and software support costs.

You can achieve this through the use of appropriate:

- network connectivity
- hardware configuration (memory, CPU speed, disk space)
- operating system software
- subject gateway database and associated software
- Web server software

## Hardware and software requirements; issues for managers

---

### Reliability

You want your subject gateway to be reliable. You want it to be available for use for as much of the time as possible - preferably 24 hours a day, 365 days a year. In order to achieve this, there are several issues you will need to think about when you are setting up and running the gateway.

#### Use reliable hardware

Use reliable hardware to run your subject gateway. This probably means using hardware with which you are familiar. Get a hardware support contract for your machine with an appropriate call-out time. If you are nervous, make sure that you can offer your service from some other hardware if your main kit is seriously broken. If you are really nervous, set aside a machine specifically for this purpose. As regards cost, you are likely to get a much better price/performance ratio by choosing Intel (PC) hardware. However, remember that you are likely to be accessing your disks heavily during subject gateway operation so choose an appropriate disk configuration and connection method.

#### Use reliable software

Remember that a subject gateway operates in a hostile networked environment and needs to support multiple users. Choose an operating system that can reliably handle this. Again, it may be sensible to choose an operating system with which you are familiar. However, it is worth noting that UNIX-based operating systems have a much longer track record of providing Internet-based services. Think carefully before choosing anything else! Much of the software developed by the DESIRE project is aimed at (or will only run under) UNIX-based operating systems. If you've chosen Intel-based hardware, using Linux as the operating system is an obvious choice. Remember that you may need software support both for your operating system and for the subject gateway software that you are running. If you prefer to pay for such support, fine; but remember that the freely available and fairly informal support which is usually available for Open Source software through mailing lists and Web sites can often be extremely good. Remember also that your subject gateway software is likely to rely on a separate Web server; the widely deployed, well maintained and supported and freely available Apache Web server is a sensible choice.

#### Make sure your data is regularly backed up

What happens when something goes seriously wrong with your machine: a disk crashes or you are hacked and your data is deleted? Make sure that all your software and data is backed up in such a way that you can quickly and easily recover your service. You may choose some sort of RAID architecture for your disks. You may choose to copy your data automatically to a second disk partition. In any case, you are advised to archive your data to tape regularly. You may even do all three of these things ... but do something! And don't forget your software and configuration files; in the event of a serious problem you may need to re-install absolutely everything!

#### Make sure your server is secure

An insecure server is a disaster waiting to happen. Follow the advice in your operating system manuals concerning security. Apply all known security patches and get someone in your team on to the right mailing lists so that you find out about potential problems early. Only run the minimum number of network services that you have to. Position your machine behind a firewall if you can, with access to the Internet only on those ports that you actually need.

#### Coping with external problems

Your subject gateway will rely on various external services. If your network connection goes down, you can't offer a service. If your DNS entry isn't available for some time, people may be unable to access you. An off-site secondary for your DNS entries is a good idea; an off-continent secondary is even better! As your subject gateway grows, you might think about mirroring your service at another location. One way of achieving this is to have a reciprocal mirroring arrangement with another subject gateway.

#### Staffing issues

Unless you hand over completely the running and administration of your subject gateway server to a third party, you are highly likely to need one technically competent member of staff to run a subject gateway. For DESIRE developed software solutions, this will mean someone familiar with administering UNIX machines. Familiarity with the Perl programming language would be a distinct

advantage as well. Other software solutions may not require UNIX or Perl experience; however, a technical understanding of the issues related to the operation of a networked service will be very helpful.

### **Responsiveness and efficiency**

### **Hardware and software issues**

More details concerning hardware and software issues are given in the Systems Requirements Specifics section. The main rules of thumb are:

- hardware requirements will be software-specific - in particular, database-specific. Check your software manual!
- more memory is likely to mean better performance
- faster CPU speed is likely to mean better performance
- Linux will give better performance than NT given the same hardware
- NT and Perl may not mix well
- more network bandwidth means better performance
- multiple DNS secondaries will give better performance

#### **CROSS REFERENCE**

[System requirements specifics, hardware and software](#)

### **Network and design issues**

The design of the Web interface to your subject gateway will have an effect on the efficiency with which you use the available network bandwidth. Make as many of your pages as possible suitable for caching. For example, most of your browsable interface (assuming that you have one) can probably be designed so that it can be cached by remote Web caches and at the Web browser. Your user interface will be much more responsive because of this.

#### **CROSS REFERENCE**

[User interface implementation](#)

### **Scalability**

Scalability is discussed in more detail in the Scalability section. As a general point it is worth noting that:

- supporting more users may require more memory and more network bandwidth
- having more records in the database may require more memory and more disk space
- introducing new service scenarios may require more memory and more disk space

#### **CROSS REFERENCE**

[Scalability](#)

### **Costs**

Unless you are very lucky, the hardware on which you run your subject gateway is going to cost money. As mentioned above, Intel-based hardware is likely to give a much better price/performance ratio than other hardware. Software may well be free - all the software developed by the DESIRE project will be made available on an Open Source basis. Hardware and software support is likely to cost money; though again it is worth noting that the support you can get for free from the Internet community may well be good enough for your needs (and may even be better than that provided commercially). Technical staff will cost money.

## Future proofing

---

Software and hardware systems need to be regularly reviewed to measure how far they are meeting business requirements. The gateway will want to choose software and hardware solutions which provide sufficient flexibility to accommodate change. Such products will probably:

- offer regular upgrades
- comply with open standards
- respond to customer requests
- impose no restrictions which tie you to that product, for example by ensuring that you have access to proprietary specifications of data structures which may be needed to convert to a new supplier's format The gateway will want to ensure that decisions regarding the choice of products are informed by strategic objectives, for example:
- use products that have a good reputation in areas which are important for the gateway (by being innovative, reliable, flexible, customisable . . . )
- use products that support inter-working with key collaborators
- implement systems with potential audiences in mind (the technologies they use, the features they value)

### EXAMPLE

#### Scout/SOSIG mirroring

SOSIG, the Social Science Information Gateway, is a ROADS database of over 5500 Internet resource descriptions operated by ILRT at the University of Bristol in the UK. In order to make the database more accessible to end-users in North America, SOSIG has been working closely with staff from the Internet Scout Project, located at the University of Wisconsin-Madison (USA) and funded by the National Science Foundation. A mutual mirroring service has been set up so that users from North America can access a mirror of SOSIG, based on the Scout server, and European users can access a mirror of Scout from the SOSIG server. The SOSIG ROADS database is mirrored weekly using some locally developed scripts that make a 'tar' copy of the complete SOSIG ROADS installation (after making some site-specific changes).

#### CROSS REFERENCE

[Co-operation between gateways](#)

## Glossary

---

**DNS** - Domain Name Server. A general-purpose distributed, replicated, data query service chiefly used on Internet for translating hostnames into Internet addresses.

**Linux** - Linux is a free Unix-type operating system originally created by Linus Torvalds with the assistance of developers around the world.

**RAID** - Redundant Arrays of Independent Disks

**ROADS** - Resource Organisation And Discovery in Subject-based Services

## References

---

Apache, <http://www.apache.org/>

Internet Scout Project - SOSIG mirror, [http://scout18.cs.wisc.edu/sosig\\_mirror/](http://scout18.cs.wisc.edu/sosig_mirror/)

Linux, <http://www.linux.org/>

SOSIG, <http://www.sosig.ac.uk/>

AE. Frisch, *Essential System Administration, 2nd ed.* (ISBN: 1-56592-127-5).  
<http://www.oreilly.com/catalog/esa2/>

B. Laurie & P. Laurie, *Apache: The Definitive Guide, 2nd ed.* (ISBN: 1-56592-528-9).  
<http://www.oreilly.com/catalog/apache2/>

M. Loukides, *System Performance Tuning* (ISBN: 0-937175-60-9).  
<http://www.oreilly.com/catalog/spt/>

E. Siever, et al., *Linux in a Nutshell: A Desktop Quick Reference* (ISBN: 1-56592-585-8).  
<http://www.oreilly.com/catalog/linuxnut2/>

## Credits

---

Chapter author: [Andy Powell](#)

## 1.5. Maintenance requirements

### In this chapter...

---

- the importance of maintenance
- estimating maintenance requirements

### Introduction

---

Information gateways need to be maintained in two key areas:

- collection management
- server integrity and functionality

Without adequate maintenance in these two areas a gateway is vulnerable to undermining its core aims and objectives; being a quality-controlled portal to online information resources. The key strength of an information gateway is in the quality of its data and the reliability of its service. Without adequate maintenance both of these areas are susceptible to developing weaknesses and problems.

## The importance of maintenance

---

### Server integrity and functionality

All Web sites and services need some degree of Web server maintenance. A competent system administrator and Webmaster can easily carry out much of this technical maintenance. Additionally many maintenance tasks can be readily automated, thereby reducing the requirements for direct human intervention. However there is still a need for someone to keep an eye on things, such as monitor system performance and deal with any day-to-day maintenance issues that may arise. Without this maintenance there is a real risk that any problems with the Web server will not be picked up until users find them. If users experience regular problems with Web sites they are likely to lose trust in the sites in question. Loss of trust often results in lost users.

Information gateways have the additional requirement that they need regular and sometimes extensive maintenance of the resource catalogue. Because the resource catalogue is at the heart of the gateway (it is the very reason why people use the gateway), then failure to maintain this aspect of the gateway can lead to serious problems in quality of service and content. Problems in this area directly effect user confidence in the gateway. Without user confidence and quality assurance gateways can rapidly lose users and fail to attract new ones.

### Collection management

Because of the dynamic nature of the Internet, a catalogue of Internet resources is going to require a certain degree of maintenance in order to keep the catalogue up to date. Online resources come and go, are available one day and not the next (the fluidity of many online documents is detailed elsewhere - Collection management). This makes collection management an important part of any gateway's maintenance requirements.

### CROSS REFERENCE

[Collection management](#)

## Estimating maintenance requirements

---

Estimating maintenance requirements for an information gateway can be a difficult task. Key factors that should be considered are:

- what is the scope of the gateway?
- how quickly is the gateway resource catalogue scheduled to grow?
- what is the perceived lifetime of the gateway?
- how heavily will the gateway be used?

Generally the larger the scope, the quicker the scheduled growth, the longer the lifetime and the more heavily used the gateway is the more maintenance will be required.

### Server integrity and functionality

Server maintenance will be largely constant regardless of the size of the gateway. If the gateway has its own dedicated server then there will be basic machine level administration tasks. If the gateway is hosted virtually (i.e. multiple Web sites on the same machine), then a large proportion of the maintenance will be shared with other sites on that machine.

For more details on hardware and software maintenance see the System requirements specifics, hardware and software chapter.

### CROSS REFERENCE

[System requirements specifics, hardware and software](#)

Virtual hosting maintenance can be as little as a few hours a week of staff effort, sometimes even less. Dedicated servers are going to require more maintenance but with the right planning and set-up the maintenance requirements can be kept below one day per week in staff effort.

These low levels of maintenance can be achieved only with careful planning and setting up of the gateway from the start. Obviously when problems arise (they do even for the best-planned gateway) maintenance requirements can be considerably more time consuming.

### Collection management

Collection management and associated maintenance requirements are closely linked to the size of the catalogue and resources database. Validating records, link checking and updating resource descriptions will be related to the number of records that are being dealt with. As the catalogue grows expect to spend 10-15% of the overall cataloguing time on collection management maintenance and related tasks.

### REMEMBER

General Web sites often require an unexpectedly high level of maintenance. It has been estimated that "as a rule of thumb, the annual maintenance budget for a website should be about the same as the initial cost of building the site, with 50 percent as an absolute minimum."

Jakob Nielsen: 1997  
<http://www.useit.com/alertbox/9706b.html>

## References

---

Jakob Nielsen *Top Ten Mistakes of Web Management. Alertbox, June 15 1997.*  
<http://www.useit.com/alertbox/9706b.html>

## Credits

---

Chapter author: [Martin Belcher](#)

## Section 2 : Information Issues (Print Version)

### Target audience

---

Section 2 of this handbook is aimed at gateway staff responsible for information management - the subject specialists and information professionals who will consider the content and organisation of the information within the gateway.

It aims to cover the important decisions that need to be made when setting up a new gateway (such as choosing a metadata format, designing a use interface, writing a selection policy) but also covers issues that arise in the day-to-day running of an existing gateway (such as cataloguing, resource discovery and publicity and promotion).

Each chapter offers some background, practical tips and hints, key references, a glossary, case studies and examples. Watch out for the  **CROSS REFERENCE** that will take you to related sections elsewhere in the handbook.

### Contents

---

Section 1 : [Strategic Issues](#)

Section 2 : Information Issues

1. [Quality selection](#)
2. [Resource discovery](#)
3. [Metadata formats](#)
4. [Cataloguing](#)
5. [Subject classification, browsing and searching](#)
6. [Collection management](#)
7. [Working with information providers](#)
8. [Publicity and promotion](#)
9. [User interface design](#)
10. [Integration of robot and manual indexes](#)
11. [Distributed cataloguing](#)
12. [Multi-lingual issues](#)
13. [Co-operation between gateways](#)

Section 3 : [Technical Issues](#)

## 2.1. Quality selection: ensuring the quality of your collection

### In this chapter...

---

- why develop and publish a selection policy for your gateway?
- creating a scope policy and selection criteria for your gateway
- guidelines for selecting and evaluating Internet resources
- skills and training required by gateway staff in selection and evaluation
- changing your selection criteria over time
- quality ratings/labelling/PICS and other Internet initiatives in this area

### Introduction

---

Subject gateways are sometimes called the Internet equivalent of a library, and in terms of the selection process this is certainly true.

Gateways are characterised by the focus and quality of their collections. They aim to provide their users with a quality controlled environment in which to search for information on the Internet and they do this by building selective collections where every resource that the gateway points to has been carefully selected for its quality.

The selection process involves people making value judgements about Internet resources and selecting only those resources that satisfy certain quality criteria.

But what constitutes a 'high quality' Internet resource? Information gateways need to use a service-driven definition of quality, where resources are selected for their relevance to the user

group as well as their inherent features.

Selecting resources for a gateway therefore requires a clear understanding of the information needs of the end-users, as well of as the pros and cons of the design features of Internet sites.

Information gateways consciously emphasise the importance of skilled human involvement in the assessment and 'quality control' of their selected Internet resources. Selection and evaluation of resources for a gateway is typically done by a librarian or subject specialist, reflecting the fact that selection is based on an evaluation of the semantic content of the resources.

A formal selection policy can support the development of a consistent and coherent collection of high quality Internet resources.

### **Why develop and publish a selection policy for your gateway?**

---

Many subject guides on the Internet do not explicitly state their selection policies, but there are a number of advantages in developing a formal selection policy for a gateway and publishing it on your site:

- it helps users to appreciate that the service is selective and quality controlled
- it helps users to understand the level of quality of information they will find when using the service
- it helps gateway staff to be consistent in their selection and to maintain the quality of the collection
- it can be used to train new staff
- it ensures consistency in collections that are developed by a distributed team

By publishing your selection policy on the gateway you can help your users to conceptualise the nature of the collection they are using. On the Web, users are very often faced with a search box or an index, and it is not always easy for them to understand exactly what they are searching. An explicit selection policy can help them to understand the nature of your gateway service. The Centre for Information Quality Management (CIQM) recommends that database providers offer a '*published specification*' or '*user-level agreement*' to '*lessen the gap between user expectations and the reality of searching*' (Armstrong, 1997). A formal selection policy can help to meet with this recommendation.

The integrity of a collection will depend on there being some consistency in the type and quality of resources that your staff decide to include in the collection. A formal selection policy can help to ensure that the selection is consistent and that the quality of the collection remains high.

A selection policy can ensure that the same member of staff makes consistent judgements about what they include in the collection. It can also ensure that different members of the staff team make consistent judgements and that they are all using the same selection criteria.

The selection policy can help new staff to understand quickly both the nature of the collection and the criteria they should use when selecting new resources to add to the gateway.

A formal policy can also help to ensure consistency of selection within a distributed team. For example, if a number of gateways are working collaboratively, an agreed selection policy can help to ensure that the combined collection has a consistent level of quality.

### **What is a selection policy?**

---

In an information environment, a selection policy defines the criteria used for selecting resources to add to a collection. It will typically outline the scope of the collection and the criteria used when new resources are selected for the collection. The scope policy relates to the needs of the target user group, while the selection criteria relate to the inherent features of the Internet resources.

#### **Defining the scope of the collection**

Subject gateways do not aim to include every resource available on the Internet. The scope of a gateway defines the boundaries of the collection. The scope policy is therefore a broad statement of the parameters of the collection.

The scope policy of a service states what is and is not to be included in the catalogue. In the



selection process, the scope of the service will affect the first decisions made about the quality of the resources. Those falling outside the scope will be rejected and the rest will have the quality criteria applied to them.

The scope criteria are the first filter through which the resources pass. They will tend to involve clear decisions; either a resource falls within the scope or it does not.

A scope statement will typically outline:

- the subject areas covered by the gateway
- the types of resources covered by the gateway

It may also outline:

- language parameters (e.g. whether the gateway only includes resources in a certain language)
- geographical parameters (e.g. whether the gateway only includes resources from a particular country)
- other parameters of relevance to the user group served

#### E X A M P L E

Examples of scope policies

- [SOSIG scope policy](#)
- [DutchESS scope policy \(in Dutch\)](#)

### Defining the quality selection criteria

Subject gateways do not generally aim to point to every Internet resource that falls within their subject area and scope. They are characterised by their quality control, aiming to point only to the best resources available for their subject area and audience.

The selection criteria outline the qualities that a resource must have to be included in the collection.

#### E X A M P L E

Examples of quality selection criteria

- [The European Link Treasury](#)
- [Evaluating Internet Resources for SOSIG](#)
- [Länkskafferiets Kvalitetskriterier \(in Swedish\)](#)
- [Scout Report selection criteria](#)
- [DutchESS \(in Dutch\)](#)
- [EELS Engineering Electronic Library, Sweden quality and selection policy](#)

### Developing a selection policy for your gateway

How should a gateway develop its selection policy? Each gateway needs to develop its own unique set of selection criteria to take the information needs of the user group and the aims of the service into account.

The first steps are to define:

1. your target user group
2. the information needs of the user group
3. the aims and objectives of the gateway (balancing what you'd like to cover with what you have the resources to cover)

Once these steps have been taken, it is a matter of defining a formal scope policy and a set of

selection criteria.

The DESIRE project has created some tools for creating a scope and selection policy. The guidelines are not prescriptive and are designed to help an institution or service develop its own tailor-made policies in the light of its aims and audience. A comprehensive list of criteria is given, from which criteria relevant to the individual service can be chosen. The list has been drawn from a 'state of the art review' of current practice, library and Web literature.

### Creating a scope policy

Some possible criteria for creating your scope policy are given below. For each heading you will need to outline the parameters to be used in your gateway. Not all of these will be appropriate for your audience and you may need to add additional criteria.

INFORMATION COVERAGE	
Subject Matter	<ul style="list-style-type: none"> <li>• what subject matter is appropriate for the target audience?</li> <li>• are there any subjects which will be censored (e.g. for ethical reasons, such as resources produced by hate groups or resources about bomb-making/paedophilia etc.)</li> <li>• how important is the subject matter of linked sites?</li> </ul>
Acceptable Types of Resource	<ul style="list-style-type: none"> <li>• what types of resource are appropriate for the target audience?</li> <li>• is the information scholarly rather than popular?</li> <li>• does the resource contain more than just a list of links?</li> <li>• is the site either proven to be or expected to be durable?</li> <li>• would a resource intended for use by an individual or local group be acceptable?</li> <li>• is it innovative - does it contain breakthrough design elements?</li> </ul>
Acceptable Sources	<ul style="list-style-type: none"> <li>• which sources of information are acceptable/appropriate for the target audience?</li> <li>• are academic, government, commercial, trade/industry, non-profit private sources all acceptable?</li> <li>• are pages maintained by individual enthusiasts (e.g. students) acceptable?</li> <li>• is biased information acceptable, and are opinions and ideologies acceptable?</li> </ul>
Acceptable Levels of Difficulty	<ul style="list-style-type: none"> <li>• which sources of information are acceptable/appropriate for the target audience?</li> <li>• are pages maintained by individual enthusiasts (e.g. students) acceptable?</li> <li>• is biased information acceptable, and are opinions and ideologies acceptable?</li> </ul>
Acceptable Levels of Difficulty	<ul style="list-style-type: none"> <li>• what level of resource is appropriate for the target audience? (e.g. users may be school children or may be academics)</li> </ul>
Advertising	<ul style="list-style-type: none"> <li>• are resources that contain advertising acceptable?</li> <li>• is there a limit to the amount of advertising that is acceptable?</li> <li>• are there any forms of advertising that will be censored?</li> </ul>
ACCESS	
Cost	<ul style="list-style-type: none"> <li>• how is charging going to affect selection - is the service only going to point to resources that are free to access?</li> <li>• are there any price limits in terms of the access charge?</li> <li>• what if resources are under copyright?</li> </ul>
Technology	<ul style="list-style-type: none"> <li>• what technologies are appropriate for the target audience? (forms, ismaps, databases, CGI scripts, Java applications, frames, etc.)</li> <li>• what connectivity does your audience have and how will this affect selection?</li> </ul>

	<ul style="list-style-type: none"> <li>• what software do your users have and how will this affect selection? (e.g. will resources that work well in graphical browsers but not in line browsers be accepted?)</li> <li>• what hardware do your users have and how will this affect selection?</li> </ul>
Registration	<ul style="list-style-type: none"> <li>• will the service accept resources where user-registration is necessary before the resource can be accessed?</li> <li>• is online registration acceptable?</li> <li>• if users must negotiate written contracts before access is possible, is this acceptable?</li> </ul>
Special Needs	<ul style="list-style-type: none"> <li>• do your users have any special needs that will affect the resources selected? (e.g. large print or audio options for disabled users)</li> </ul>
<b>METADATA AND CATALOGUING ISSUES</b>	
Granularity	<ul style="list-style-type: none"> <li>• at what level will resources be selected/catalogued?</li> <li>• will resources be considered at the Web site/Usenet group level or the Web page/Usenet article level?</li> </ul>
Resource description	<ul style="list-style-type: none"> <li>• what is the minimum amount of information needed to create a resource description in your catalogue, i.e. what basic information <b>MUST</b> a resource contain to be selected? (e.g. in a WWW document, contact details, last update details, etc.)</li> <li>• is there sufficient information to create a descriptive record?</li> <li>• will the service accept resources with/without specific metadata?</li> </ul>
<b>GEOGRAPHICAL ISSUES</b>	
Geographical Restraints	<ul style="list-style-type: none"> <li>• are any geographical restraints appropriate for your audience?</li> <li>• will the service cover information produced locally, from particular countries, particular continents or worldwide?</li> </ul>
Language	<ul style="list-style-type: none"> <li>• in which languages are resources acceptable/appropriate to your target audience?</li> </ul>

### Creating quality selection criteria

Once you have defined the scope of your gateway, you will need to outline the level of quality that is acceptable within each individual resource.

A list of possible quality selection criteria is given below, from which criteria relevant to the individual service can be picked.

#### Content criteria: evaluating the information

- validity
- authority and reputation of source
- accuracy
- comprehensiveness
- uniqueness
- composition and organisation
- currency, adequacy of maintenance

#### Form criteria: evaluating the medium

- ease of navigation
- provision of user support
- use of recognised standards
- appropriate use of technology
- aesthetics

### Process criteria: evaluating the system

- information integrity (work of the information provider)
- site integrity (work of the Webmaster/site manager)
- system integrity (work of the systems administrator)

Fuller description of each of these criteria and examples can be found in an online tutorial called 'Internet Detective':



#### [Internet Detective](#)

Internet Detective is an interactive, online tutorial which provides an introduction to the issues of information quality on the Internet and teaches the skills required to evaluate critically the quality of an Internet resource. There is no charge, it takes around two hours to complete and it has interactive quizzes and exercises to lighten the learning process.

#### [Selection criteria for quality controlled information gateways](#)

This is a lengthy, peer-reviewed report which describes the DESIRE research into the development of quality systems and selection criteria for subject gateways. This report will be of interest to people wishing to see the research and methodology that lay behind the development of the lists of criteria given above. The lists resulted from a 'state of the art' review of quality issues, both within subject gateways and in other sectors, notably the private sector and industry.

### Guidelines for selecting and evaluating Internet resources

The staff responsible for selecting new resources to add to the gateway will need to be able to select resources that together create a consistent and coherent collection of high quality Internet resources.

What constitutes a 'high quality' Internet resource? The definition of quality used here has been drawn from the commercial sector, where quality is seen to be closely related to customer satisfaction and to developing systems of continuous improvement. In the context of a subject gateway, the quality of a resource will depend on the users of the service, and the nature of the service, as well as the internal features of the resource itself. We suggest that for information gateways *'a high quality Internet resource is one that meets the information needs of the user'*.

This is a service-oriented definition, and so, when evaluating the quality of Internet resources, gateway staff must consider the user group that they are serving as much as the Internet resources they are evaluating.

SOSIG (The Social Science Information Gateway) has come up with five steps that describe the selection process for gateway staff:

#### EXAMPLE

##### **SOSIG selection procedure: Five steps to quality control**

##### **Before you start - get to know the quality of SOSIG**

- read the SOSIG scope policy, which outlines the subjects and types of resources that are acceptable
- become familiar with the SOSIG service, especially the coverage of the collection; browse the database to see the kinds of resources that are acceptable
- become familiar with the SOSIG quality selection criteria outlined in these Web pages

##### **Finding resources**

You may find it easier to divide the selection process into two stages:

1. Spend time finding resources on the Internet and bookmarking those with potential.
2. Go back to the bookmark list later to spend time evaluating each resource in some detail.

Once you have found a resource to evaluate, there are five steps to quality control, which are summarised below.

### **1. Ensure that the resource falls within the scope of SOSIG**

This is the most important filter through which all resources should pass - if it isn't relevant then reject it! You can use the scope policy for guidance. Most important of all is to ensure that the resource is social science related! You can look at the browsing pages to see which subject areas the service covers.

### **2. Search the SOSIG collection**

To avoid duplication within the SOSIG collection, it is essential that you go to 'Search SOSIG' and check that the resource is not already in the database. Consider how the resource will add to the SOSIG collection (this will get easier the more you get to know SOSIG). The coverage and balance of the collection is important. Try to find resources for subject areas that are not well covered.

### **3. Evaluate the content of the information**

Content criteria are based on the information the resources actually contain. Of the criteria relating to the resources themselves, the content criteria are the most important. Content criteria should take precedence over form criteria - SOSIG users are likely to care more about getting the information that they need than about the form it takes.

### **4. Evaluate the form of the information**

Form criteria relate to the medium, design and presentation of the resource. Some evaluation of the form can be made by considering the ease of navigation, provision of user support, and design. Resources should rarely be rejected on design points alone, but there may be factors which should be mentioned in your description of the resource (e.g. if a resource comes in a form that some users will not be able to access).

### **5. Evaluate the processes set up to support the resource**

Process criteria relate to the fact that Internet resources can be volatile and can lack integrity. Some evaluation of the processes set up to support a resource is necessary. These may involve personnel as well as computer systems. You need to evaluate the likelihood that a resource will be adequately maintained over time and that it will remain current and stable.

Quality resources can now be added to SOSIG via the WWW catalogue form

## **Skills and training required by gateway staff in selection and evaluation**

The choices made by the staff who select resources for a gateway will determine the nature of the collection. Recruitment and training of staff will therefore be a critical choice for your gateway.

### **Recruiting staff**

Subject gateways typically employ librarians or subject specialists to select Internet resources to add to the gateways. This reflects an acceptance that to build a high quality collection you need:

- a good understanding of the information needs of your target user group
- to base selection on semantic judgements about the relevance and value of resources to your users
- to have knowledge and expertise in the subject
- to have knowledge and experience of information resources
- skills in critical evaluation of information resources

Recruiting skilled and knowledgeable staff will help ensure the integrity of the gateway collection.

### **Training staff**

Staff will need to be consistent in their selection criteria if the collection is to develop consistently. They will need to be familiar with the scope and selection criteria of your gateway, but will also need to develop skills for evaluating Internet resources. Training staff may involve:

- 'editorial meetings'- where all the selection staff discuss the criteria to be used
- creating a staff manual - giving staff paper or online copies of the selection policy
- developing exercises and examples based on Web sites to evaluate
- asking staff to complete the 'Internet Detective' online tutorial
- monitoring the sites selected by new staff to check they comply with the selection policy
- setting up an email list for all staff to discuss and debate any quality issues that arise

### **Changing your selection criteria over time**

---

It may be necessary to update a selection policy, as the priorities for selection may change over time as a gateway collection matures.

#### **Adapting scope policies**

A new gateway may wish to focus on developing a core collection very quickly before broadening the parameters. The scope may be much narrower in the early stages of collection development. For example, a new gateway may set narrow parameters for things such as:

- granularity (e.g. focus on Web sites as opposed to Web pages)
- subjects covered (e.g. prioritise generic resources over resources for very rarely researched subjects)
- geographic boundaries (e.g. focus on UK resources before adding those from elsewhere)
- types of resource (e.g. focus on Web sites as opposed to mailing lists or newsgroups)

A more mature gateway on the other hand may broaden its scope once a core collection has been developed to include resources beyond the very narrow scope initially used. It may choose to extend its subject coverage, work at a finer level of granularity or include resources from different countries and of different types. These decisions should be reflected in the scope policy of the service.

#### **Adapting selection criteria**

The Internet offers uneven coverage of subjects, and this may affect the quality selection criteria used within different parts of a gateway collection.

For example, if a subject comes within the scope of the gateway but very few resources can be found about that subject, it may be that less stringent quality criteria should be used, to ensure that there is at least some subject coverage.

Conversely, if there are many resources available for a subject, then very stringent quality criteria may be used to ensure that the highest quality resources are selected in preference to others with the same subject coverage.

These issues relate to collection management, which is discussed in the [Collection Management](#) chapter of this handbook.

### **Quality ratings/labelling/PICS and other initiatives in this area**

---

The Web and metadata communities have been exploring the potential for automated approaches to quality-related aspects of information management on the Internet. The main aim has been to create a system where the quality of an Internet resource can be described in a machine-readable form. If this were to be achieved a number of scenarios would become possible. For example:

- search engines could retrieve or rank resources according to aspects of their quality
- users could search for resources using particular quality requirements (e.g. only peer reviewed journals, or resources that work with version 3.1 of Netscape, or resources that have been approved by a librarian)
- users could recommend and rate Internet resources in a standard format and share these ratings

There have been two main challenges:

1. Creating the technological infrastructure to support machine-readable quality ratings.
2. Creating metadata vocabularies to describe various quality attributes of Internet resources.

### **PICS and RDF**

PICS and RDF both aim to provide a technological infrastructure to support machine-readable quality ratings.

PICS stands for Platform for Internet Content Selection. It has been approved by the W3C (World Wide Web Consortium) as an agreed standard for associating labels (metadata) with Web sites or Web pages. Essentially, these labels refer to the information content of the sites, and therefore provide a means of recording information about aspects of their quality. PICS has most famously been used to support the development of services that aim to protect children from X-rated sites on the Internet.

RDF stands for Resource Description Framework and is a standard approved by the W3C. It has emerged as a successor to PICS, offering a broader infrastructure for assigning metadata labels to Internet sites and pages. RDF can be used with many different metadata vocabularies, and certainly there is potential for it to be used with a vocabulary that describes the quality of an Internet resource.

### **Metadata vocabularies for quality**

The second challenge has been to create metadata vocabularies to describe various quality attributes of Internet resources. At the time of writing no vocabulary has emerged but work is under way, particularly within the medical community, to create metadata labels for quality that can be incorporated into Internet resource discovery services.

With the basic RDF framework in place, it is now possible for different communities to create their own quality vocabularies and apply them to their own services.

### **How does this work relate to Information gateways?**

This work has the potential to offer gateways a number of interesting possibilities, for example:

- Internet cataloguers may use quality ratings to help them find high quality resources to add to their gateway
- gateways may create machine-readable quality labels
- they may incorporate user ratings into their services

The missing link, as things stand, is the development of quality vocabularies. Gateways may see it as their role to create such vocabularies and to use RDF to create machine-readable metadata about the quality of Internet resources. At present we cannot offer an example of a gateway doing this, but some key sites where new developments will appear are listed below.

### **EXAMPLE**

#### **Examples of recent work with PICS and quality ratings**

- [RDF Home Page](#)
- [PICS Home Page](#)
- [Quality Ratings in an RDF Environment](#)

## Glossary

---

**DutchESS** Dutch Electronic Subject Service  
**EELS** Engineering Electronic Library Sweden  
**PICS** Platform for Internet Content Selection  
**RDF** Resource Description Framework  
**SOSIG** Social Science Information Gateway

## References

---

DutchESS, <http://www.konbib.nl/dutchess/>

EELS, <http://www.ub.lu.se/eel/>

European Link Treasury, <http://www.en.eun.org/news/european-link-treasury.html>

Information Quality WWW Virtual Library, <http://www.ciolek.com/WWWVL-InfoQuality.html>

Internet Detective, <http://www.sosig.ac.uk/desire/internet-detective.html>

Länkskafferiet (Link Larder), <http://länkskafferiet.skolverket.se/information/kvalitetskriterier.html>

PICS Home Page, <http://www.w3.org/PICS/>

RDF Home Page, <http://www.w3.org/RDF/>

Scout Report, <http://scout.cs.wisc.edu/index.html>

SOSIG, <http://www.sosig.ac.uk/>

J. Alexander & M. A. Tate, *Evaluating Web Resources*,  
<http://www2.widener.edu/Wolfgang-Memorial-Library/webeval.htm>

C. Armstrong, 'Metadata, PICS and Quality', *Ariadne Issue 9*. 1997  
<http://www.ariadne.ac.uk/issue9/pics/>

N. Auer, *Bibliography on Evaluating Internet Resources*  
<http://www.lib.vt.edu/research/libinst/evalbiblio.html>

D. Brickley, T. Gardner, R. Heery & D. Hiom, *Recommendations on Implementation of Quality Ratings in an RDF Environment*.  
<http://www.desire.org/html/research/deliverables/D3.2/>

A. Cooke, *Finding Quality on the Internet: a guide for librarians and information professionals*, (London: Library Association Publishing, 1999. ISBN: 1-85604-267-7).

## Credits

---

Chapter author: [Emma Place](#)

With contributions from: Michael Day, Debra Hiom, Ann-Sofie Zettergren



## 2.2. Resource discovery

### In this chapter...

---

- the resource discovery process - ensuring new Internet resources are found to add to your gateway
- systems for gateway managers - to support efficient resource discovery within your team
- strategies for gateway staff - to continuously locate high quality resources on the Internet
- case studies - resource discovery tips and hints from existing gateways
- new and mature gateways - different resource discovery issues for different gateways

### Introduction

---

Subject gateways should aim to describe the best resources that the Internet has to offer in their field and for their target audience. They need to:

- point to the highest quality networked resources currently available
- point to new networked resources as they appear

Finding high quality resources on the Internet can be a time-consuming job - which of course, is exactly why gateways exist - to save the end-user some of the time and commitment required to discover and retrieve high quality information on the Internet.

Locating resources to add to your gateway will require one of the biggest investments of staff time and effort, and so it is important to find efficient and effective methods of working at this task:

- gateway managers need to ensure that systems to support resource discovery are in place
- individual gateway staff need to develop their own strategies for locating as many high quality resources as efficiently as possible

### Resource discovery issues for gateway managers

---

Gateway managers will need to provide the systems and strategies to support efficient resource discovery within their team.

Resource discovery is labour-intensive and efficient strategies can help to maximise the number of resources added to the gateway. This section suggests some of the systems that managers can put in place to support efficient resource discovery within the team:

1. Avoid duplicated effort.
2. Find the right people for the job.
3. Provide training in resource discovery.
4. Set up support systems for resource discovery staff.
5. Set up systems to encourage your user community to suggest resources.

#### 1. Avoiding duplicated effort

Duplicated effort can be wasted effort. There are issues of duplication:

- between gateways
- within the team

#### Avoid duplication with other gateways

It is worth finding out whether other gateways already describe Internet resources in your field. If there are other gateways you have to ask yourself whether it really makes sense to spend time and effort cataloguing the same resources twice. If existing gateways are already describing resources relevant to your users you should consider:

- collaboration with other gateways (to avoid cataloguing the same resources twice)
- cross-searching your gateway with other gateways so that your users can search more than one simultaneously

- sharing metadata records

### CROSS REFERENCE

[Co-operation between gateways](#)

#### **Avoid duplication within your team**

Time can be wasted if members of your team are all trawling the same sources. Consider developing a team strategy for resource discovery. For example by:

- giving people different subject responsibilities - so they are each hunting for resources in a different discipline
- giving people different monitoring responsibilities - so they are each monitoring different sources (email lists/URLs/current awareness services etc.)

#### **E X A M P L E**

##### **Example of a team dividing resource discovery responsibilities**

SOSIG has divided responsibilities among the team of core staff and section editors as follows:

**Section Editors:** each have responsibility for a particular SUBJECT area

**Central staff:** have responsibility for trawling generic sources and for monitoring suggestions of sites sent in by users

See: <http://www.sosig.ac.uk/contact.html>

## **2. Find the right people for the job**

It will be financial and political considerations which determine whom you can take on to do the job of resource discovery, as with recruiting staff for cataloguing.

### CROSS REFERENCE

[Subject indexing and classification, Distributed cataloguing](#)

#### **Volunteers?**

*Pros:* may be cheap and plentiful

*Cons:* may be inconsistent and unreliable in their contribution and it may be difficult to find volunteers with the subject expertise to select the high quality resources you want

#### **Subject specialists?**

*Pros:* may know of the best sources to use to discover relevant resources for your gateway and should be able to assess resources effectively, given their subject knowledge.

*Cons:* may be expensive, short of time, difficult to recruit and unable or unwilling to spend time cataloguing

#### **Librarians/information professionals?**

*Pros:* have training in selecting resources to meet the information needs of users and also may be able to catalogue resources in addition to selecting them, since they may have training in cataloguing/information retrieval issues.

*Cons:* may be expensive/difficult to recruit

**REMEMBER**

- Internet skills can be taught more easily than subject expertise!
- Librarians may be more willing and able to catalogue resources than to discover them

**3. Provide training in resource discovery**

The Internet is always growing and changing, so there are always new tips and hints to be learned in Internet resource discovery - training staff can improve skills and effectiveness. Training may include:

- offering lists of sources for staff to use
- offering demonstrations and hands-on work with different resource discovery tools
- brainstorming ideas within the team to share resource discovery strategies

**4. Set up support systems for resource discovery staff**

The following are ideas for support systems for resource discovery staff:

- create Web documents that list resource discovery strategies appropriate to your gateway
- set up a mailing list for resource discovery staff so that the team can share knowledge of any useful new sources or techniques they find - and so they can talk about issues that arise
- set up meetings for resource discovery staff to share stories of successful and unsuccessful strategies which they have found.

**EXAMPLE****Example of a support system for gateway staff**

1. SOSIG has created a Web page for section editors, which lists possible resource strategies: ['Finding Internet resources for SOSIG: strategies and sources'](#)
2. A mailing list has been set up for section editors to share news of any new, effective strategies they discover.
3. Twice a year the section editors come together and compare experiences of the most effective and the most ineffective (!) resource discovery strategies.

**5. Set up systems to encourage your user community to suggest resources**

Why not let the resources come to you! Encourage your users to send you details of any sites which they think should be added to the gateway. You will need:

1. to publicise an email address or Web form for submissions
2. to publicise your scope and selection criteria

**CROSS REFERENCE**

[Quality selection](#)

**TIPS**

- Web forms are great because they encourage users to generate the appropriate metadata - and they may have good ideas about keywords and descriptions
- make sure your selection criteria are freely available, to try to discourage inappropriate resources from being submitted and to make it clear that not all submissions will be accepted
- a quick thank-you message to users is good PR and can encourage them to submit again. If you are getting a lot of submissions - create a standard

- *submit again* if you are getting a lot of submissions - create a standard courtesy reply
- publicise the fact that you welcome submissions from your user community. If you run an email list associated with your gateway, (\*\*CROSS REFERENCE publicity and promotion) you can send out occasional reminders to subscribers

#### E X A M P L E

#### Examples of Web forms for users to submit resources

- [DutchESS](#)
- [EEVL](#)
- [SOSIG](#)

### Resource Discovery Strategies for Staff

Gateway staff do the 'leg work' for SOSIG users - joining the lists, monitoring the sites and doing the searches that many users do not have the time to do, filtering out items that are of poor quality or irrelevant to the users.

It's easy to waste time when surfing the Internet - gateway staff need to develop efficient and effective strategies for locating high quality Internet resources. Some strategies are suggested below.

#### Resource discovery tools and methods

1. Browsing strategies
2. Mailing lists and their archives
3. Distribution lists and current awareness services
4. Search tools
5. Newsgroups and discussion forums
6. URL-minders and Web agents
7. Non-Internet sources

#### 1. Browsing strategies

One of the richest sources of resources will be existing Web pages - especially authoritative ones in your field which list related or recommended resources. Trawling these sites is the equivalent of citation pearl-growing or snowballing, traditionally done by researchers looking for references - if they find one useful resource, they will follow the references from that resource to find others.

#### Trawling home pages of known experts

If you know of experts in your field, do a search to see if they have their own Web page. You may find that:

1. They have published their work on the Web.
2. They have collected a list of links (and, given their knowledge and expertise, they will be worth checking out!)

Bookmark any that look as if they may be developed over time, so that you can check them again in the future.

#### Trawling organisational home pages

Many organisations now have their own Web sites. These can be useful in two ways:

1. They may include primary resources for you to catalogue.
2. They may have lists of links selected by people with subject knowledge which you could trawl.

Consider which organisations are relevant to your audience and try to keep in touch with developments concerning them.



Take time to do a search for the most relevant organisational sites for you and organise them in a bookmark folder, so you can take a look at them periodically. Only bookmark the best - you won't have time to trawl too many.

If you are creating a gateway for an academic audience then it can pay to monitor university Web pages. Look for:

- library Web sites - as many librarians are now building collections of Internet links
- academic departments' Web sites - where lecturers and researchers may publish their work or may create lists of links

#### EXAMPLE

**Examples of some starting points useful for academic gateways:**

- [College and University Home Pages \(world-wide\)](#) - alphabetical listing
- [EUNI - List of European Universities](#)
- [Library and Related Sources](#) (includes a list of libraries worldwide)

#### Trawling subject-based sites

Many sites have a section of 'links' which can be mined for new resources. The better quality the original site, the better the related links are likely to be:

- find the most important sites in your field and look at all the links they recommend
- look for 'What's New' or 'Latest News' features on trusted sites
- bookmark these link pages or 'What's New' pages to check regularly, or consider putting the URLs into a Web Agent or URL-minder (see below) so that they can let you know when anything new is added

#### EXAMPLE

**Examples of the types of pages that could be bookmarked or monitored by a minder/agent:**

- ['What's New' on Europe](#) - the Web server of the European Union
- [NewJour](#): Recent Issues

## 2. Mailing lists and their archives

Joining and monitoring email lists/checking mailing list archives

People often use email lists to announce new resources they have made available on the Internet.

You have two possible strategies here:

1. Joining the lists and reading messages via your email
2. Bookmarking the Web archives of the lists (if they have them) and making periodic checks on them



Don't join so many lists that your own email becomes unmanageable. If you can, filter your email so that messages from lists don't get mixed up with all your other mail. For very busy email lists it is probably more time-effective to make a regular scan of the archives. Set up a bookmark file for 'Archives to Check Regularly'

### Subject-based lists

If you can find a list that is relevant to your subject area and audience, you have a rich source. In the early days it's worth doing a search for relevant lists and asking colleagues to recommend them.

#### EXAMPLE

##### Examples of sites which can help you to find mailing lists

- [Liszt](#) - Directory of email groups and discussion lists  
A directory of email groups and discussion lists, including listserv, listproc, majordomo and Mailbase lists. Also offers a directory of newsgroups. The search facility makes this a quick way of finding lists on a particular subject.
- [Mailbase](#) - The UK's major electronic mailing list service
- [The Directory of Scholarly and Professional E-Conferences](#) - A directory designed to list the Internet communication groups and services likely to be of interest to academics and professionals.

### Generic email lists that announce new Internet sites

A number of email lists exist to alert people to new Internet sites. Be warned - these lists can be prolific!

### 3. Distribution lists and current awareness services

Internet current awareness services come in different forms and are becoming more sophisticated. Free email subscription services will send you updates, bulletins and email publications on a regular basis. It may be worth subscribing to services that are run by key individuals or organisations in your subject area. Other services are emerging where you can create your own personal profile on the Web, which the service then uses to email you incoming information that is likely to interest you.

#### EXAMPLE

##### Examples of current awareness services

- [What's New in WWW Social Sciences Online Newsletter](#) - users can subscribe to receive emails listing of new or improved WWW sites
- [Internet Resources Newsletter](#) - A free, monthly, non-subscription newsletter for academics, students, engineers, scientists & social scientists. ISSN: 1361-9381

### 4. Search tools

Searching the Internet can be time-consuming, since many of the search tools retrieve huge numbers of hits which take a lot of time to work through. However, searching can be a good strategy in some cases:

- targeted searching, i.e. looking for a specific resource
- building up a specific section of your collection

In our experience, search engines can be a waste of time if broad search terms such as 'social psychology' are used. Highly focused searching based on known sources, however, can be fruitful. For example, if you have a list of well-respected journals or organisations in your field, you could search for them by name, to see whether they have a presence on the Internet. A number of hints for finding the leads for focused searching are recommended:

- use other sources, e.g. directories, to find things to search for
- use a subject-specific site to get lists of dates/organisations/names to search on
- search for Internet equivalents of printed materials, e.g. scholarly journals or academic publishers

- search for specific dates or people
- search for important organisations to see if they are publishing anything of value on the Internet
- use leads from your knowledge of the field

### Search Engines

These are good for finding LOTS of information and for finding very precise pieces of information (so if you know exactly what you're after they can be very effective).



Get to know how to use one search engine very well, rather than lots of them very badly. Take time to read the Help pages for the search engine and learn how to use the Advanced Search options.

Be aware that search engines change over time and that different ones are more effective for searching for different types of information - do some research to find the best one for your needs.

Bookmark complex searches so that you can run them again periodically to see if anything new has appeared.

#### EXAMPLE

##### Examples of ways to find out about Internet search tools

- [Search Tools](#) - a list constructed by Manchester Metropolitan University's Department of Information and Communications
- [Search Engine Corner \(a regular column in Ariadne\)](#)
- [Search Engine Watch](#)

### 5. Newsgroups and discussion forums

Internet discussion forums are a powerful and fun way to communicate with people around the world who are interested in the same things as you. Thanks to the Internet's rapid growth and the exploding popularity of the World Wide Web, people from all walks of life now participate on a regular basis.

#### EXAMPLE

##### Example of a source for Newsgroups

[DejaNews](#) offers access to tens of thousands of Usenet groups and discussion forums. It can help you to find those forums relevant to your user groups, but it may also be worth following a few yourself to see if any other Internet resources are talked about that would be appropriate for your gateway.

### 6. URL-minders and Web agents

Some free Web services exist that help you to monitor changes made to Internet resources or to inform you of new sites that might interest you. You register the URLs of the sites you wish to monitor or search queries you would like to have done and the service sends you an email whenever a change is made to these resources or the search yields new results.

#### EXAMPLE

##### Examples of URL-minders and Web agents

- [NetMind's Mind-it](#)
- [The Informant](#)

Remember that these are automated services and will not always yield high quality results.



- Remember that the more URLs you register, the more email you will get - so don't set up more than you can cope with! If you can, set up email filters to separate these messages from the rest of your mail.

## 7. Non-Internet sources

You don't have to use the Internet to learn about Internet sites. Consider using non-Internet sources:

- **talk to people** - your users/experts in your field/Internet enthusiasts and get their recommended sites
- **look at the bookmarks** of these people if they publish them on the Web - if not, then ask them to let you get access to them another way
- **scan printed publications** e.g. specialist journals, newspapers, newsletters, magazines
- **watch out for URLs** - which are increasingly appearing everywhere from billboards to TV to the side of cornflake packets!



### REMEMBER

It's chaos out there so don't expect resource discovery to be without its problems:

- expect information overload and develop systems to manage it effectively
- let serendipity play a role
- be open to adopting new strategies and changing your old ways - the Internet is always changing
- be open minded - take the Alexander Fleming attitude - there may be millions of petri dishes containing nothing more than a load of jelly, but keep your wits about you - what looks like a mould may turn out to be penicillin!

## Issues for new gateways

---

New gateways may have different priorities for resource discovery from mature gateways as they will be focussing on developing a core collection very quickly. New gateways may want to consider the following issues:

- target efforts to make sure that you include the most important resources first
- balance the collection to ensure you have at least a few resources for all the subject areas you cover
- divide responsibilities among your team
- don't duplicate other gateways
- be absolutely clear of your scope and selection criteria before you start the resource discovery process

## Issues for mature gateways

---

Mature gateways will have already developed a core collection and may have widened their scope. Staff will need to adjust their resource discovery strategies in line with this. Mature gateways may consider the following issues:

- collection management - you need to ensure that all the different subject areas within your collection are growing at the same rate - target efforts at areas that are falling behind and require development.
- ensure that all areas of the collection are comparable in quality
- focus on strategies for finding new resources AS THEY APPEAR
- build your community - to encourage more submissions from users and information providers



 **CROSS REFERENCE**

[Quality selection: Changing your selection criteria over time](#)

## Glossary

---

**DutchESS** Dutch Electronic Subject Service  
**EEVL** Edinburgh Engineering Virtual Library  
**EUNI** List of European Universities, provided by Adminet in France  
**SOSIG** Social Science Information Gateway  
**URL-minder** a service based in California, USA, twhich enables you to track changes made to Web sites and URLs

## References

---

- College and University Home Pages (world-wide),  
<http://www.rirr.cnuce.cnr.it/universities/univ.html>
- Dejanews, <http://www.dejanews.com/>
- The Directory of Scholarly and Professional E-Conferences, <http://www.n2h2.com/KOVACS/>
- DutchESS, <http://www.konbib.nl/dutchess/>
- EEVL, <http://www.eevl.ac.uk/>
- EUNI - List of European Universities, [http://www.ensmp.fr/~scherer/euni/euni\\_list.html](http://www.ensmp.fr/~scherer/euni/euni_list.html)
- The Informant, <http://informant.dartmouth.edu/>
- Library and Related Sources, <http://www.exeter.ac.uk/~ijtised/lib/wwwlibs.html>
- Liszt, <http://www.liszt.com/>
- Mailbase, <http://www.mailbase.ac.uk/>
- Mind-it, <http://mindit.netmind.com/>
- NewJour: Recent Issues, <http://gort.ucsd.edu/newjour/nj2/>
- Search Engine Corner, <http://www.ariadne.ac.uk/issue19/search-engines/>
- Search Engine Watch, <http://searchenginewatch.com/>
- Manchester Metropolitan University's Department of Information and Communications Search Tools, <http://www.mmu.ac.uk/h-ss/dic/main/search.htm>
- The Social Science Research Grapevine, <http://www.grapevine.bris.ac.uk/>
- SOSIG, <http://www.sosig.ac.uk>
- What's New in WWW Social Sciences Online Newsletter, <http://www.mmu.ac.uk/h-ss/dic/main/search.htm>
- 'What's New' on the Web server of the European Union,  
<http://europa.eu.int/geninfo/whatsnew.htm>
- A. S. McNab & I. R. Winship, *How to find out about new resources on the Internet*, The New Review of Information Networking (1995), 147-53.

Association of Public Data Users and International Association for Social Science Information Service and Technology (IASSIST), *Strategies for Searching for Information on the Internet*. [http://dpls.dacc.wisc.edu/www\\_searchers.html](http://dpls.dacc.wisc.edu/www_searchers.html)

TERENA & M. Isaacs, *Internet Users' Guide to Network Resource Tools*, Addison Wesley Longman: 1998

E. Worsfold, *Finding Internet resources for SOSIG - strategies and sources*, 1997  
<http://sosig.ac.uk/desire/esig.html>

## Credits

---

Chapter author: [Emma Place](#)

With contributions from: Lisa Gray (OMNI), Debra Hiom (SOSIG), Linda Kerr (EEVL), John Kirriemuir (OMNI), Roddy McLeaod (EEVL), Kate Sharp (Biz/ed)

## 2.3. Metadata formats

### In this chapter...

---

- why create metadata records?
- types of metadata attributes
- standard metadata formats
- choosing metadata attributes and formats for your gateway
- format conversion and future proofing

### Introduction

---

Information gateways are characterised by their creation of third-party metadata records - individual descriptions of Internet resources held in a database that have separate fields for different attributes of the resources, such as title, author, URL etc. These resource descriptions are used to:

- help users learn more about the Internet resources (from a trusted third-party)
- support information search and retrieval

Gateways adopt the approach where metadata is created by a third party ie. an independent subject specialist or information professional, rather than the creator of the resource. This enables the quality control for which gateways are renowned - the resource descriptions all assume a standard format and are generated manually (at least in part) to enable high quality metadata that benefits for semantic judgements about the nature and origin of the resources.

The metadata created by gateways is their greatest asset - adding value to the Internet resources by creating independent, standardised third-party descriptions.

The decision of which metadata format to use is an important one as it impacts on the searching capabilities of the gateway and the value of the descriptions to the end-users. The creation of metadata will be one of the most time-consuming tasks in running a gateway and so a balance between value and cost may be required in deciding on a format.

This chapter will introduce some of these issues and provide some background information that information gateway managers will need to consider when choosing a metadata format for their gateway.

## Why create metadata records?

---

Information gateways are services that give access to networked resources in particular subject areas, linguistic domains, and so on. Many Internet portals simply comprise of sets of Web pages with lists of hyperlinks on a static Web page, perhaps with annotations, however, this approach has distinct disadvantages:

- the portal can be browsed, but with no database it cannot be searched effectively
- maintaining the portal is time consuming as all edits and additions require manual changes to the HTML

Gateways take advantage of database technologies which gets over both these problems, but requires that a standard format be used for creating and storing the resource descriptions. Metadata formats are structured formats for Internet resource descriptions. For gateways, the metadata formats are the forms or templates that need to be filled in by the cataloguers to create a resource description.

The use of metadata by an information gateway has many benefits over the simple HTML list approach, for example:

- the metadata has structure and so can form the basis of far more advanced search facilities within a gateway (e.g. fielded searching, such as searching by title or author)
- the metadata can be converted to other formats or be otherwise persuaded to interoperate with different search and retrieve protocols
- it is easier to maintain a database of resource descriptions than a large number of HTML files. Administrative metadata can also be used to record when resources need to be re-evaluated or removed from the database

## Metadata attributes

---

Gateways staff will need to agree on the attributes of an Internet resource that they wish to describe. Metadata can be grouped into various kinds according to their use within the gateway. They might include:

### Descriptive

Descriptive metadata contain information which may be usefully returned from a search of the gateway. A user may be able to decide from this information whether it is worth spending time looking at the resource itself.

- title
- short title (e.g. an acronym of the full title)
- alternative title (e.g. title of resource in another language)
- subtitle
- description
- URI (or other location)
- author
- language
- character set encoding
- organisation - either creating or hosting the resource-
- medium (e.g. text/images/audio/video)
- type of resource (using types appropriate to your gateway)
- physical medium
- copyright owner
- availability (is payment or registration needed?)
- software required for access (e.g. specific browsers, MIDI software)
- quality rating
- intended audience (e.g. undergraduate level)

### Subject

Subject metadata can facilitate effective searching. They can also be used to organise the browsing structure of your gateway. A fuller discussion can be found in the



### Subject indexing and classification

- keywords
- classification code
- classification system - must accompany classification code!
- terms from thesauri
- subject headings

### **Administrative**

Administrative metadata are intended primarily to assist the gateway staff in maintaining the gateway. They are of less concern to users and may not be visible to them; however, they can be used, for example, to check that resource descriptions are still current.

- resource maintainer
- date of addition of resource to gateway
- date record was last updated
- date resource was last changed
- review-by date
- expiry date (e.g. of a conference announcement)
- submitter of resource
- cataloguer of resource
- origin of record (if gateway has collaborators)
- rights ownership

### **EXAMPLE**

ROADS templates contain relatively simple administrative metadata attributes like the following:

To-Be-Reviewed-Date:  
 Record-Last-Verified-Email:  
 Record-Last-Verified-Date:  
 Comments:  
 Record-Last-Modified-Date:  
 Record-Last-Modified-Email:  
 Record-Created-Date:  
 Record-Created-Email:

Consideration of which particular administrative functions are required and an assessment of which particular administrative metadata elements are needed will be an important part of choosing (or adapting) a metadata format for use in a particular information gateway.

### **Core metadata**

The possible metadata fields listed above are by no means exhaustive, but including them all would require considerable effort both in initial cataloguing and in keeping records up to date. Not all of them might be appropriate to your gateway.

Attempts have been made to define standards for a 'core' of metadata which should be regarded as a bare minimum. One such standard is the Dublin Core.

### **EXAMPLE**

**Dublin Core currently involves 15 core elements:**

1. Title
2. Author or Creator
3. Subject and Keywords
4. Description
5. Publisher
6. Other Contributor
7. Date
8. Resource Type
9. Format
10. Resource Identifier
11. Source
12. Language

12. Language
13. Relation
14. Coverage
15. Rights Management

[http://purl.oclc.org/dc/about/element\\_set.htm](http://purl.oclc.org/dc/about/element_set.htm)

ROADS offers a number of metadata templates designed for different types of Internet resources. Each template contains attributes specific to the type of Internet resource. For example, the template for describing a mailarchive will have a different set of fields from the template for describing a Web document. ROADS also maintains a 'template registry' where the metadata fields used in the various kinds of ROADS templates are recorded. This ensures that ROADS services are potentially interoperable in this area. New fields can be nominated for addition to the registry.

#### E X A M P L E

ROADS offers metadata formats for the following types of Internet resource:

ROADS template-types:

COLLECTION - experimental  
 DATASET  
 DOCUMENT  
 DUBLINCORE  
 EVENT - experimental  
 IMAGE  
 MAILARCHIVE  
 PROJECT  
 SERVICE  
 SOFTWARE  
 SOUND  
 TRAINING MATERIALS  
 USENET  
 VIDEO

<http://www.ukoln.ac.uk/metadata/roads/templates/>

### Choosing metadata attributes

---

You should think carefully about which metadata attributes your gateway is going to use, and their format, when you first set up the gateway. If you do not, you may find yourself constrained by the absence of useful metadata, or have to add a new metadata field or convert an existing field to a different format when you already have several thousand resources in your database. Moreover, decisions about metadata will in turn affect the design of your interface (especially the parts of it used for cataloguing and/or submitting new resources for consideration).

#### CROSS REFERENCE

[Cataloguing](#)

#### **Which metadata fields could be usefully searched on by your users?**

You should consider your potential user community and also the nature of the resources which your gateway will cover. For example, if your gateway is intended to cover only geographically local resources in one language, a 'language' field will not be very informative unless your gateway is going to be cross-searched with others elsewhere.

#### **And how are they going to search them?**

This will affect not only what metadata fields you provide but also the cataloguing rules you adopt. For example, if you are ranking searches by the frequency of the occurrence of the search term, you may wish to make descriptions similar in length, otherwise resources with long descriptions may be more likely to be returned high up the order.

#### CROSS REFERENCE

[Subject indexing and classification](#)

**Which metadata fields will be displayed to the users of the gateway?**

Will they need to be converted from the form in which they are stored and if so does an easy way of converting them exist?

**Which metadata fields will be used for housekeeping by the gateway staff and how?**

Metadata can supply information for partially automating this otherwise laborious aspect of gateway management. For example, you can have an automatic email sent to maintainers of resources occasionally to ask whether they have made any changes, or set a web-page tracking tool to monitor changes to resources.

 **CROSS REFERENCE**

[Collection management](#)

**Which if any are optional?**

If you are collaborating (or thinking of it), which metadata fields will be shared with your collaborators? Are they likely to want extra information, such as language, which you would not otherwise include in your metadata? You will need to use the same schemes for e.g. classification or have a usable crosswalk to convert between schemes. You should also think about the issue of copyright.

 **CROSS REFERENCE**

[Co-operation between gateways, Interoperability](#)

**Are you going to display your metadata in the same format as that in which you store it?**

If not, you will need a way of converting between formats.

**Can any of the software you are using generate useful metadata?**

For example, ROADS automatically records when a template was last updated. You may wish to use in addition software for creating metadata (see below). Harvesting software, if used, may also be able to harvest metadata.

 **CROSS REFERENCE**

[Harvesting, indexing and automated metadata collection](#)

**Who will generate metadata fields (and which ones?).**

Metadata may be supplied by:

- information providers
- gateway users
- cataloguers for the gateway
- subject editors for the gateway
- core gateway staff
- another gateway working in collaboration with you
- automatic generation by software

How much cross-checking will there be? (Time will need to be allowed for this).

**If you are allowing gateway users or information providers to submit resources, what information should they supply?**

What information may they also supply optionally? How important is it that (for example) descriptions or keywords are consistent across the gateway? If this is important, can you supply cataloguing rules or other guidance to help information providers and others who are submitting resources? How much effort can be expended on editing their contributions, given that gateway

users and information providers cannot be compelled to follow your cataloguing rules?

#### CROSS REFERENCE

[Working with information providers](#)

How might you ensure that information such as dates is in a consistent format? Possible methods include:

- pulldown menus on forms
- authority files
- cataloguing rules

#### CROSS REFERENCE

[Cataloguing](#)

**In what language are your metadata records going to be kept?**

If this is different from the language of some of your resources, are you going to make any provision for searching in that language (e.g. an 'alternative title' field)?

#### CROSS REFERENCE

[Multi-lingual issues](#)

## Standard metadata formats

Information gateway managers will need to make decisions about which metadata format (or formats) to use within their service at a very early stage of its development. At present, however, the existence of a large and varied range of metadata formats and initiatives complicates these decisions.

It is worth remembering also that the choice of metadata formats will often be influenced by other factors, both technological and social. For example, an information gateway that wishes to use the ROADS software toolkit with little modification will currently need to use the ROADS template format, or something very similar to it. Again, where gateway cross-searching or interoperability is seen to be important, there may be technical reasons why one format may have advantages over another.

The nature of metadata development means that at any one time there are likely to be a variety of formats that could be chosen as the basis of an information gateway. For example, a review of metadata formats undertaken under DESIRE I identified and described over twenty formats that were in use (or under development) in 1996 (Dempsey et al., 1997). In order to help analyse the different metadata formats described in the review, the DESIRE I study produced a typology of metadata based upon their underlying complexity.

Band One	Band Two	Band Three	
[simple]	-----	-----	[complex]
(full text indexes)	(simple structured generic formats)	(more complex structure, domain specific)	(part of a larger semantic framework)
Proprietary formats	Proprietary formats Dublin Core ROADS templates LDIF	FGDC MARC	TEI headers EAD CIMI

Figure 1. Typology of metadata formats (adapted from Dempsey and Heery, 1998).

## Choosing a metadata format

Choosing a format from the variety of existing ones will depend upon various factors. In general, current information gateways tend to use relatively simple generic formats with some structure ('Band Two' formats such as ROADS templates or Dublin Core). These formats have the twin advantages of simplicity, which means that they are relatively easy to create and maintain, and the existence of some structure, which facilitates both interoperability and format conversion. However, in particular circumstances there may be good arguments for basing an information gateway on more complex formats ('Band Three' formats such as MARC or TEI headers) if this offers some competitive advantage to the gateway. For example, the USMARC format has been used for the cataloguing of Internet resources in the InterCat project and it would be possible to set up MARC-based information gateways. However, the use of these more complex formats may have implications for the level of expertise (technical and other) that would be required for cataloguing and may have other costs.

As noted before, the choice of a particular format may be dictated by technological or social factors. For example, particular gateway software may dictate the use (or non-use) of particular formats. Information gateways that, for example, are running the ROADS software without much modification will need either to use one of the existing templates defined by the ROADS project or to create new (and similar) templates in the form of attribute-value pairs.

### Example format 1: Dublin Core

The Dublin Core (DC) is the result of an international and interdisciplinary initiative to define a core set of metadata elements for electronic resources, primarily for resource discovery on the Internet. DC was initially conceived as a simple format that could be used for author-generated descriptions of Web resources. However, the format has also attracted the attention of resource description professionals from a variety of communities such as libraries, museums, archives and government agencies.

#### EXAMPLE

##### Example of a DC based gateway

EdNA (Education Network Australia):

EdNA - an information gateway for Australian education resources - uses a metadata standard that is based on the DC element set. The owners of documents are encouraged to embed metadata within their documents where it can be read by the EdNA resource harvester and transferred to the EdNA database.

- EdNA: <http://www.edna.edu.au/EdNA/>

The format has been developed by means of a series of invitational workshops, the first being held in Dublin, Ohio in March 1995. The workshop series and related work has resulted in the definition of fifteen core metadata elements as RFC 2413 (Weibel et al., 1998). These elements are intended to be repeatable and extensible in any application.

The initial focus of DC was the Web, so the initiative has concentrated on the production of draft guidance for the encoding of DC elements, first in HTML (Kunze, 1999) and more recently in XML/RDF (e.g. Miller, Miller and Brickley, 1999).

#### EXAMPLE

##### Example of DC metadata embedded in HTML

```
<link rel="schema.DC" href="http://purl.org/dc">
<meta name="DC.Title" content="Southampton Oceanography Centre (SOC)">
<meta name="DC.Creator" content="Bruce Dupee (b.dupee@soc.soton.ac.uk)">
<meta name="DC.Subject" content="oceanography, marine, technology, geology, seafloor,
education, science, research, ships, vessels">
<meta name="DC.Description" content="An introduction to the services provided by the
Southampton
Oceanography Centre - a joint venture between the University of Southampton and the Natural
Environment
Research Council. Includes information on internal departments and divisions, and the National
Oceanographic Library">
<meta name="DC.Publisher" content="NERC Computer Services">
```



```
<meta name="DC.Identifier" content="NERC Computer Services" />
<meta name="DC.Date" scheme="WTN8601" content="1999-06-08">
<meta name="DC.Type" content="Text">
<meta name="DC.Format" content="text/html">
<meta name="DC.Format" content="7985 bytes">
<meta name="DC.Identifier" content="http://www.soc.soton.ac.uk/">
```

Metadata created by DC-dot, a service that will retrieve a Web page and automatically generate Dublin Core metadata, either as HTML <META> tags or as RDF/XML, suitable for embedding in the page header.

- DC-dot: <http://www.ukoln.ac.uk/cgi-bin/dcdot.pl>
- Dublin Core: <http://purl.oclc.org/dc>

### Example format 2: ROADS templates

ROADS templates are a development of the IAFA templates originally developed for anonymous FTP archives (Deutsch et al., 1994). IAFA templates are a simple text-based metadata format consisting of predefined sets of attribute-value pairs. Templates exist for a number of different resource types, but the templates most commonly used in existing ROADS-based gateways are those designated SERVICE, DOCUMENT and MAILARCHIVE.

#### E X A M P L E

##### Example of part of a ROADS SERVICE template

```
Template-Type: SERVICE
Handle: 840738289-29226
Title: Southampton Oceanography Centre
URI-v1: http://www.soc.soton.ac.uk/
Admin-Email-v1: webmaster@mail.soc.soton.ac.uk
Publisher-Name-v1: University of Southampton
Publisher-Postal-v1: Southampton Oceanography Centre, University of Southampton, Waterfront
Campus, European Way, Southampton SO14 3ZH, United Kingdom
Publisher-City-v1: Southampton
Publisher-Country-v1: UK
Publisher-Phone-v1: +44 (0)1703 596666
Description: An introduction to the services provided by the Southampton Oceanography Centre
- a joint venture between the University of Southampton and the Natural Environment Research
Council. Includes information on internal departments and divisions, and the National
Oceanographic Library
Keywords: Southampton Oceanography Centre; Natural Environment Research Council; NERC;
Subject-Descriptor-v1: 551.46
Subject-Descriptor-Scheme-v1: DDC21
Record-Last-Modified-Date: Wed, 12 May 1999 18:24:49 +0000
Record-Last-Modified-Email: cataloguer@subject-gateway.ac.uk
Record-Created-Date: Wed, 12 May 1999 18:24:49 +0000
Record-Created-Email: cataloguer@subject-gateway.ac.uk
```

- ROADS project: <http://www.ilt.bris.ac.uk/roads/>

### Format conversion

One of the advantages of using well-defined and structured metadata formats is that this allows conversion into other formats when necessary. This is useful in two main circumstances:

1. When a gateway wants to change to using a different metadata format. For example, a gateway that currently uses a custom-built database management system with a Web interface might want to run the ROADS software to take advantage of cross-searching facilities. The gateway's existing records would therefore need to be converted into ROADS templates. These types of conversion will be required periodically as information gateway software and its associated metadata evolve.
2. To aid interoperability.

Format conversion is facilitated by the creation of crosswalks (or mapping tables) between metadata formats. Crosswalks can be used as the basis for the production of a specific conversion

program or for the production of search systems that would permit the interrogation of heterogeneous metadata formats. A number of metadata format crosswalks have been published. One of the earliest DC-based crosswalks mapped Dublin Core to USMARC (Caplan and Guenther, 1996) and other crosswalks exist for other formats including Text Encoding Initiative (TEI) headers, ROADS templates and a variety of MARC formats, including the Universal MARC format (UNIMARC). A collection of metadata mappings is maintained on the UKOLN Web site (Day, 1996).

### CROSS REFERENCE

[Interoperability](#)

#### EXAMPLE

##### Examples of metadata conversion projects

###### Nordic Metadata Project

The Nordic Metadata Project produced a variety of tools designed to aid the wider utilisation of Dublin Core (Hakala et al., 1998). The toolkit included a utility called d2m, a Dublin Core to MARC converter that converts Dublin Core metadata embedded in HTML into a variety of Nordic MARC formats and USMARC.

- d2m: <http://www.bibsys.no/meta/d2m/>

###### BIBLINK project

The BIBLINK project developed a custom-built software system (the BIBLINK Workspace) which converts metadata produced by publishers into the UNIMARC format for use by participating national bibliographic agencies (Day, Heery and Powell, 1999). The UNIMARC records can in turn be converted into other formats (usually MARC-based) used by these national bibliographic agencies, who can then enhance them for inclusion in their national bibliography and (possibly) for returning this enhanced record to the publisher. The metadata conversion process in the BIBLINK Workspace uses metadata crosswalks produced for the project by UKOLN (e.g. Day, 1998a).

- BIBLINK: <http://hosted.ukoln.ac.uk/biblink/>

### Future proofing

Any choices concerning metadata will need to take into account possible future developments. The gateway may decide to expand by including new types of descriptions (possibly for new types of resource such as images or multimedia) or to include additional metadata (such as descriptions aimed at alternative audiences, rights metadata, digital preservation data). At the simplest level, updates and extensions to existing metadata element sets need to be accommodated. The gateway may want to ensure that:

- metadata creation tools can be easily extended to deal with new elements and new formats
- the system has sufficient flexibility to allow a variety of formats to be imported and exported

Within the lifetime of the gateway, it may have to migrate to a different system which will require different metadata formats, whether these are new versions of existing formats or completely different. Re-structuring the metadata can be done more efficiently if the gateway follows some general guidelines for the content of metadata. Such guidelines might include recommendations that:

- metadata formats and rules for content are agreed among collaborating gateways (this means that gateways can share costs of converting their data)
- gateways implement local usages by means of local processing rather than by incorporating them into the data (for example, adding punctuation and other presentational enhancements by software processing rather than by storing it as part of the data)
- there are as few local variants to standard metadata formats as possible. (For example, variant element names can be displayed using local processing rather than by storing non-standard element names.)
- collaborate with other gateways so that migration can take advantage of economies of scale.

## Conclusions

---

Choosing a metadata format is one of the most important decisions that needs to be made when setting up an information gateway. It is vital that the format is able to work with the software that forms the basis of the gateway service and it should also contain all fields (including administrative metadata) that have been identified as appropriate for the service in question (or the format should be extensible). It is possible that ongoing changes in technologies may require periodic conversion of the gateway database into new formats. This process will require the production of metadata crosswalks and/or format conversion programs.

## References

---

BIBLINK, <http://hosted.ukoln.ac.uk/biblink/>

d2m, <http://www.bibsys.no/meta/d2m/>

DC-dot, <http://www.ukoln.ac.uk/cgi-bin/dcdot.pl>

Dublin Core, <http://purl.oclc.org/dc>

EdNA, <http://www.edna.edu.au/EdNA/>

InterCat, <http://purl.org/net/intercat>

ROADS, <http://www.ilrt.bris.ac.uk/roads/>

P. L. Caplan & R. S. Guenther, 'Metadata for Internet resources: the Dublin Core Metadata Element Set and its mapping to USMARC', *Cataloging and Classification Quarterly* 22 (3/4) (1996), 43-58.

M. Day, *Interoperability between metadata formats* (Bath: UKOLN, 1996).  
<http://www.ukoln.ac.uk/metadata/interoperability/>

M. Day, *Mapping BIBLINK Core (BC) to UNIMARC. BIBLINK project document* (Bath: UKOLN, 10 September 1998).  
<http://hosted.ukoln.ac.uk/biblink/wp10/bc-unimarc.html>

M. Day, R. Heery & A. Powell, 'National bibliographic records in the digital information environment: metadata, links and standards', *Journal of Documentation* 55 (1) (1999), 16-32.

L. Demspey & R. Heery, 'Metadata: a current view of practice and issues', *Journal of Documentation* 54 (2) (1998), 145-172.

L. Demspey, R. Heery, M. Hamilton, D. Hiom, J. Knight, T. Koch, M. Peereboom & A. Powell, *A review of metadata: a survey of current resource description formats* (DESIRE deliverable D3.2 (1), March 1997).  
<http://www.ukoln.ac.uk/metadata/desire/overview/>

P. Deutsch, A. Emtage, M. Koster & M. Stumpf, *Publishing information on the Internet with Anonymous FTP* (Internet-Draft, September 1994).  
<http://info.webcrawler.com/mak/projects/iafa/iafa.txt>

J. Hakala, P. Hansen, O. Husby, T. Koch & S. Thorborg, *The Nordic Metadata Project: final report* (Helsinki: Helsinki University Library, July 1998).  
<http://linnea.helsinki.fi/meta/nmfinal.htm>

R. Heery, 'Review of metadata formats', *Program* 30 (4) (1996), 345-373.

R. Iannella & D. Campbell, *The A-Core: metadata about content metadata* (Internet-Draft, 21 June 1999).  
<http://metadata.net/admin/draft-iannella-admin-01.txt>

J. Kunze, *Encoding Dublin Core Metadata in HTML (Internet-Draft, 25 May 1999)*.  
<http://www.ietf.org/internet-drafts/draft-kunze-dhtml-01.txt>

O. Lassila & R. Swick, eds., *Resource Description Framework (RDF) model and syntax specification (W3C Working Draft, 1999)*.  
<http://www.w3.org/TR/WD-rdf-syntax/>

Making of America project, *The Making of America II testbed project white paper (Version 1.03, March 16 1998)*.  
[http://sunsite.berkeley.edu/MOA2/wp-v1\\_03.html](http://sunsite.berkeley.edu/MOA2/wp-v1_03.html)

E. Miller, P. Miller & D. Brickley, eds., *Guidance on expressing the Dublin Core within the Resource Description Framework (RDF) (Dublin Core Metadata Initiative, Draft Proposal, 1999)*.  
<http://www.ukoln.ac.uk/metadata/resources/dc/datamodel/WD-dc-rdf/>

S. Weibel, J. Kunze, C. Lagoze & M. Wolf, *RFC 2413, Dublin Core metadata for resource discovery (Internet Engineering Task Force, Network Working Group, September 1998)*.  
<ftp://ftp.isi.edu/in-notes/rfc2413.txt>

S. Weibel, 'The State of the Dublin Core Metadata Initiative', *D-Lib Magazine* 5 (4) (April 1999).  
<http://www.dlib.org/dlib/april99/04weibel.html>

S. L. Weibel & C. Lagoze, 'An element set to support resource discovery: the state of the Dublin Core', *International Journal on Digital Libraries*, 1(2) (January 1997), 176-186.

## Credits

---

Chapter author: [Michael Day](#)

With contributions from: Rachel Heery, Emma Place and Virginia Knight

## 2.4. Cataloguing

### In this chapter...

---

- describing Internet resources: cataloguing and metadata approaches
- metadata formats and content rules
- types of information needed by an information gateway
- developing cataloguing guidelines for a gateway
- cataloguing interfaces and maintenance

### Introduction

---

The role of cataloguing rules or guidelines is to specify how the content of a metadata format is entered. Once a metadata format has been chosen, consideration should then be given to how this metadata should be entered into the information gateway database and a set of cataloguing rules prepared.

One of the key roles of Internet subject gateways is the creation of descriptive metadata about networked resources which can be used as a basis for searching and browsing the gateway. These descriptions can also help gateway users to identify whether the resources are really what they need, potentially saving them a considerable amount of time browsing through the limited amounts of information available elsewhere on the Internet (Sha, 1995, p. 467). Therefore, one of the most important (and time-consuming) activities for a subject gateway will be the provision of these descriptions. This is the activity generally known as 'cataloguing' and is one of the key tasks of any information gateway.

## Background

Cataloguing can be defined as the creation of surrogate records which can be used to facilitate the identification, location, access and use of resources (Levy, 1995). These descriptions are usually created in accordance with certain standards (cataloguing rules and metadata formats) and will often include additional features such as classification, subject analysis and authority control (Dillon and Jul, 1996, p. 198, Bryant 1980). These tools and standards were originally developed for the cataloguing and indexing of traditional - mostly printed - collections. However, many of them have been revised to take account of resources based on newer technologies. Recent developments include:

1. ISBD(ER). In 1997, the IFLA Universal Bibliographic Control and International MARC Programme (UBCIM) published a revision of ISBD(CF) for 'Computer Files' for both online and offline 'Electronic Resources' (ISBD(ER), 1997; Sandberg-Fox and Byrum, 1998).

### EXAMPLE

#### Web page description according to ISBD(ER)

Southampton Oceanography Centre [Electronic resource]. - Electronic interactive multimedia. -- [Southampton] : University of Southampton, Southampton Oceanography Centre, cop. 199?.  
Mode of access: World Wide Web. URL: <http://www.soc.soton.ac.uk/>.  
Title from title screen.

Summary: An introduction to the services provided by the Southampton Oceanography Centre - a joint venture between the University of Southampton and the Natural Environment Research Council. Includes information on internal departments and divisions, and the National Oceanographic Library.

2. USMARC 856 field - 'Electronic Location and Access'. The use of this field enables the encoding of enough information to locate and retrieve networked resources, including an URL (Network Development and MARC Standards Office, 1997). Field 856 has been implemented in other 'flavours' of MARC such as UNIMARC (Holt, 1998).

The use of the MARC formats for describing Internet resources has been extensively tested in North America, particularly through the work of a series of OCLC projects.

### EXAMPLE

#### OCLC Internet projects

The OCLC Internet Resources project (1991-92), which resulted in the proposal for the USMARC 856 field (Dillon, et al., 1994).

- The OCLC Internet Cataloging (InterCat) project (1994-96) to test the use of the USMARC format (including the 856 field) and AACR2 cataloguing rules for describing Internet resources.
- InterCat: <http://purl.org/net/intercat> The Cooperative Online Resource Catalog (CORC) project (1998-). The project is exploring the co-operative creation and sharing of metadata by libraries. At the centre of CORC will be a catalogue containing Internet resource descriptions from a variety of sources. The project is also investigating automated methods for subject assignment, authority control and the conversion of metadata formats.
- CORC: <http://www.oclc.org/oclc/research/projects/corc/index.htm>

Information gateways build upon these practices, but have a particular focus on developing cataloguing practices and technologies that are designed specifically to manage Internet resources, taking into account the unique features of these resources.

Gateways tend to opt for more flexible and less formal cataloguing solutions, using less complex metadata formats like Dublin Core. This is largely because these formats can be flexible and quick to respond to new developments in the ever-changing Internet environment. It also helps gateways to cope with the volatility of Internet resources - one of the key challenges in Internet cataloguing - as resources change, their associated records become out of date and require frequent updating. Information gateways have sought to develop relatively simple technologies and cataloguing procedures, which provide adequate descriptions but which also support the high level of maintenance that is required.

As Clifford Lynch (1997, p. 44) has commented, if the Internet is to continue to thrive as a new means of communication, 'something very much like traditional library services will be needed to organize, access and preserve networked information'. This article also comments that combining 'the skills of the librarian and the computer scientist may help organize the anarchy of the Internet'.

## Cataloguing issues for information gateways

---

Information gateways, like libraries, need tools that facilitate the identification, location, access and use of resources; they have therefore developed (or adapted) tools that can be used for the descriptive cataloguing of Internet resources and their indexing. In this, information gateways have the distinct advantage that they can build upon the past century and a half of experience which libraries and other organisations have of the task of cataloguing. Information gateways need to work on the following:

- metadata formats
- types of descriptive information required
- content standards and cataloguing rules
- cataloguing tools and interfaces
- catalogue maintenance

### Metadata formats

Firstly, it must be noted that cataloguing issues are to some extent related to the decisions that information gateways need to make about metadata formats.

#### CROSS REFERENCE

#### [Metadata formats](#)

That said, the use of a particular metadata format does not necessarily determine the adoption of any particular description standard or set of cataloguing rules. Formats such as Dublin Core, MARC or ROADS templates are merely frameworks into which data can be entered and by which it can be retrieved. The role of cataloguing rules or guidelines is to specify how the content of this format is entered. For this reason, once a metadata format has been chosen, consideration should then be given to how this metadata should be entered into the information gateway database and a set of cataloguing rules prepared.

### Types of descriptive information required by an information gateway

During the cataloguing process for an information gateway, a resource will first be identified and selected and then described in some standardised way. Typically, a description will record a variety of different types of information:

1. Bibliographic-type descriptive information. This should include information primarily taken from the resource itself, including its title, its location (usually a URL) and the persons and organisations responsible for its content.
2. Subject information. This would include any terms added from subject schemes, such as classification codes, terms from thesauri and subject heading lists as well as any keywords added by a cataloguer. More information can be found in the chapter on Classification.
3. Administrative metadata. This includes any other information that may be useful to the management of the subject gateway. This may include information on individuals who selected or catalogued a given resource, the date that a catalogue record was created (or updated) and the dates when selected resources need to be reviewed.

### Choosing content standards and developing cataloguing rules

Once a metadata format has been adopted and decisions have been taken on the particular information that resource descriptions need to contain, it is time to start the preparation of cataloguing rules or guidelines. Such guidelines can be as detailed (or not) as a particular gateway requires. In most cases, there will not be a requirement to develop rules as comprehensive as those in AACR2, for example, but cataloguing guidelines should often contain the following things:

- a list of all possible data elements
- a brief explanation of what particular information each element is supposed to hold
- an explanation of how information should be entered into this element (the rule)

- some guidelines on the use of formats for dates, language codes, etc.
- notes of (and links to) external standards used, e.g. classification schemes, name authorities

Once developed, these guidelines can be distributed to those people who will be responsible for providing resource descriptions for the gateway.

#### E X A M P L E

##### **ROADS Cataloguing Guidelines**

The ROADS project has developed some cataloguing guidelines for the two most commonly used ROADS template types (SERVICE and DOCUMENT) which can be used as a framework for the development of cataloguing rules for new or existing information gateways (Day, 1998). These guidelines were adapted from existing practice (notably from guidelines developed by ADAM (Bradshaw, 1997) and SOSIG) and could be used as the basis for other gateways, whether based on ROADS tools or not.

- <http://www.ukoln.ac.uk/metadata/roads/cataloguing/>

Many of the decisions that need to be made relate to the particular formats that need to be used for things like dates, language codes or names.

##### **Date formats**

Dates tend to be important parts of content metadata. As well as being used to record the time when a resource was created or last modified, dates are also used to record administrative data about the metadata itself. For this reason, dates need to be entered in some agreed format so that they can be automatically processed by software. The main date formats currently in use are ISO 8601:1988 - as recommended for use in Dublin Core descriptions (Wolf and Wicksteed, 1996) - and the modified RFC 822 format used by ROADS templates (Deutsch, et al., 1994, p. 14):

- ISO 8601:1988:  
1998-06-01
- RFC 822 (as modified by RFC 1123):  
01 Jun 1998 12:00:00 GMT

##### **Language codes**

Resource descriptions tend to include an element recording the language of the intellectual content of a resource. Gateways could (and some do) record these by using the names of languages in full, e.g.:

- Language: Portuguese
- Language: Deutsch

However, natural language may not be the best way of recording this information. It would be difficult (if not impossible) for machines to be able to tell that, for example, the words 'Welsh' and 'Cymraeg' refer to the same language, or that the terms 'English' and 'Old English' refer to quite different ones. For these reasons, a number of standardised language codes have been proposed, usually based on either two or three letters (e.g. ISO 639-1:1988, RFC 1766). The best current candidate for language codes is the three-letter (known as 'Alpha-3') code ISO 639-2:1998 with more than 460 codes (Byrum, 1999):

- ISO 639-2:1998  
Language: eng  
Language: emn

##### **Name formats and authority files**

Names are one of the more problematic areas for information gateway cataloguing rules to make decisions about content. There are (in general) two main ways in which personal names can be ordered:

- Direct order:  
Author-Name-v1: Conrad Russell  
Author-Name-v1: R. Po-chia Hsia
- Inverted order:  
Author-Name-v1: Russell, Conrad  
Author-Name-v1: Hsia, R. Po-chia

However, there are a number of variations that exist within each of these ways. There is a need for rules that deal with things like titles, pseudonyms and hyphenation. These can be extremely complex. Rules concerning 'headings for persons' in AACR2 (1988 rev.), for example, take up 54 pages. Similar rules for corporate bodies take up 41 pages. In addition, in some cases there will be a requirement to be able to distinguish between two persons (or organisations) with the same name. Rules like AACR2 usually achieve this by adding more information to the name itself, e.g. dates of birth and death and titles, with appropriate punctuation:

Author-Name-v1: Hsia, R. Po-chia, 1955-  
Author-Name-v1: Newman, J. H. (John Henry), 1801-1890  
Admin-Name-v1: University of Southampton

Libraries have considerable experience of dealing with names in catalogues, as can be attested by the extremely full treatment of name entries in codes such as AACR2. The sharing of bibliographic records between institutions has additionally led to the foundation of authoritative lists of names (i.e. verified access points) with cross-references, known as name authority files.

A number of name authority lists exist, mostly produced by national libraries or national bibliographic agencies, for example:

- Library of Congress Name Authority File (LCNAF) - used by the majority of US libraries
- British Library Name Authority File - originally created for the British National Bibliography (BNB) but also now used in the British Library's own catalogues
- German-based name authority files include the Gemeinsame Körperschaftsdatei (GKD) for corporate body names and the Personennamendatei (PND) for personal names (Münnich, 1996)

At the present time name authority data tends to be national in origin, based on a variety of national formats and made available in a wide variety of ways, not always in electronic form. As one response to this problem, the AUTHOR project, funded by the Commission of the European Communities (DG XIII) as part of Computerised Bibliographic Record Actions (CoBRA), has investigated the feasibility of the international exchange of name authority data (Zillhardt and Bourdon, 1998).

If information gateways want to implement name authorities, the most logical place to start would be with the relevant national file, possibly supplemented by reference to LCNAF.

Authority files can also be used for things like geographical names or subjects. Indeed, the Library of Congress Subject Headings (LCSH) are probably the best example of a library-originated subject authority file.

### **Subject information**

Subject information, in the form of keywords, classification scheme codes, subject heading terms and so on, forms an important part of the resource descriptions provided by information gateways. Subject information can form the basis of part of the search system, or - in the case of classification codes or terms from a subject hierarchy - can form part of the gateway's browse structure. As Vizine-Goetz (1998, p. 93) has said, the 'knowledge structures that form traditional classification schemes hold great potential for improving resource description and discovery on the Internet and for organising electronic document collections'. More information on these issues can be found in the chapter on Classification.

Any cataloguing guidelines developed for information gateways need to contain information on the selected (or adapted) subject schemes and documentation will be required so that terms from these schemes can be added at the cataloguing stage. This may require reference to the published scheme itself or a link to the selected part being implemented. So, for example, a gateway based on a limited implementation of the 21st edition of the Dewey Decimal Classification (DDC21) will need at least a list of all of the classification codes in use and their meaning. More detailed implementations may require the use of the published DDC21 manuals and the employment of



suitably trained staff.

#### CROSS REFERENCE

[Subject indexing and classification](#)

### Cataloguing tools and interfaces

The creation of Internet resource descriptions for information gateways will largely take place via an interface or cataloguing tool. With some metadata formats it may be possible to create resource descriptions using text editors (e.g. for ROADS templates) or Web based tools (e.g. DC-dot for Dublin Core in HTML and RDF).

Ideally, however, information gateways need cataloguing interfaces that can be adapted for their particular needs, which contain, for example, the subject schemes used by that particular gateway as its default and including some help in the form of cataloguing rules and examples. In principle, it should be possible to embed most of the cataloguing rules developed for an information gateway inside the cataloguing interface. It should also be able to automatically validate certain elements (e.g. language codes or dates) before adding records to the database and to add certain administrative metadata.

Developing a catalogue interface, however, is a time-consuming and specialised task which is influenced by the choice of underlying software tools and metadata formats. The ROADS toolkit, for example, comes with a template editor which can be used for creating resource descriptions but this would in most cases require some customisation by the addition of guidelines for the use of subject schemes and other guidelines. Other metadata formats may have their own creation tools; for example, most MARC formats could be created using a proprietary library-based cataloguing interface.

#### CROSS REFERENCE

[User interface implementation](#)

### Catalogue maintenance

Another important factor that needs to be considered is the ongoing maintenance of the information gateway database. One of the characteristics of Internet information is that it is subject to rapid (and unadvertised) change. The content of Web pages can be frequently updated (not always for the better), their virtual locations (usually in the form of URLs) can change, and even IP addresses can expire or move to another - sometimes inappropriate - organisation. For these reasons, a considerable task for any information gateway is keeping its resource descriptions up to date. This will, in part, require the use of automated tools like link-checkers, but may also entail some periodic checking of information content (possibly based on 'expiry-date' administrative metadata or random sampling). In any case, resource descriptions will need to be periodically updated (or removed) and any cataloguing tools will need to facilitate this.

#### CROSS REFERENCE

[Collection management](#)

## Conclusions

---

As we have seen, the creation and maintenance of resource descriptions (or cataloguing) is an important part of the role of any information gateway. Gateways, therefore, need to consider in detail any cataloguing requirements that they have. This will mean decisions being made on:

- content standards - these need to be developed, whether based on Internet cataloguing guidelines such as those produced by the ROADS project or on implementations of existing standard descriptive standards like ISBD(ER)
- subject schemes - important for any browse interface to the gateway and for subject searching
- cataloguing interfaces - to ease the creation of surrogate records by gateway staff or others
- database maintenance issues - to ensure that the gateway's database is as up to date as possible

All of these decisions (and their associated activity) will require the input of specialised staff and considerable commitment in terms of time to produce (or adapt) some cataloguing guidelines, to implement a suitable cataloguing interface and to train those people who will carry out the

cataloguing task itself. Of course, there are a growing number of gateways with experience of doing these things, so new gateways would be advised to build on this experience before developing new solutions.

## Glossary

---

**AACR2** - Anglo American Cataloguing Rules, 2nd edition  
**ADAM** - Art, Design, Architecture & Media information gateway  
**BNB** - British National Bibliography  
**CoBRA** - Computerised Bibliographic Record Actions  
**CORC** - OCLC Cooperative Online Resource Catalog project  
**DDC21** - Dewey Decimal Classification, 21st edition  
**GKD** - Gemeinsame Körperschaftsdatei  
**IFLA** - International Federation of Library Associations and Institutions  
**InterCat** - OCLC Internet Cataloging project  
**ISBD** - International Standard Bibliographic Description  
**ISBD(CF)** - International Standard Bibliographic Description for Computer Files  
**ISBD(ER)** - International Standard Bibliographic Description for Electronic Resources  
**ISO** - International Standards Organisation  
**LCNAF** - Library of Congress Name Authority File  
**LCSH** - Library of Congress Subject Headings  
**MARC** - Machine-Readable Cataloguing  
**OCLC** - Online Computer Library Center  
**PND** - Personennamendatei  
**RDF** - Resource Description Framework  
**RFC** - IETF Request for Comments  
**ROADS** - Resource Organisation and Discovery in Subject-based services  
**SOSIG** - Social Science Information Gateway  
**UBCIM** - IFLA Universal Bibliographic Control and International MARC Programme  
**UNIMARC** - Universal MARC format

## References

---

CORC, <http://www.oclc.org/oclc/research/projects/corc/index.htm>

InterCat, <http://purl.org/net/intercat>

H. Alvestrand, *RFC 1766, Tags for the identification of languages (Internet Engineering Task Force, Network Working Group, March 1995).*

<ftp://ftp.isi.edu/in-notes/rfc1766.txt>

R. Braden, ed., *RFC 1123, Requirements for Internet hosts - application and support (Internet Engineering Task Force, Network Working Group, October 1989).*

<ftp://ftp.isi.edu/in-notes/rfc1123.txt>

R. Bradshaw, *Cataloguing rules for the ADAM database: a procedural manual (ADAM, the Art, Design, Architecture & Media Information Gateway, 1997).*

<http://www.adam.ac.uk/adam/reports/cat/>

P. Bryant, 'Progress in documentation: the catalogue', *Journal of Documentation* 36 (2) (1980), 133-163.

J. D. Byrum, 'ISO 639-1 and ISO 639-2: international standards for language codes. ISO 15924: international standard for names of scripts', *65th IFLA Council and General Conference, Bangkok, Thailand, 20-28 August 1999.*

<http://www.ifla.org/IV/ifla65/papers/099-155e.htm>

D. H. Crocker (rev.), *RFC 822, Standard for the format of ARPA Internet text messages (Internet Engineering Task Force, 13 August 1982).*

<ftp://ftp.isi.edu/in-notes/rfc822.txt>

M. Day, *ROADS cataloguing guidelines (Bath: UKOLN The UK Office for Library and Information Networking, 1998).*

<http://www.ukoln.ac.uk/roads/cataloguing/cataloguing-rules.html>

- P. Deutsch, A. Emtage, M. Koster & M. Stumpf, *Publishing information on the Internet with Anonymous FTP (Internet Engineering Task Force Internet Draft, September 1994)*.  
<http://info.webcrawler.com/mak/projects/iafa/iafa.txt>
- M. Dillon & E. Jul, 'Cataloging Internet resources: the convergence of libraries and Internet resources', *Cataloging & Classification Quarterly* 22 (3/4) (1996), 197-238.
- M. Dillon, E. Jul, M. Burge & C. Hickey, 'The OCLC Internet Resources Project: Toward Providing Library Services for Computer-Mediated Communication' in A. P. Bishop (ed.), *Emerging communities: integrated networked information into library services (Urbana-Champaign, Ill.: University of Illinois at Urbana Champaign, Graduate School of Library and Information Science, 1994)*, 54-69.
- M. Gorman & P. W. Winkler (ed.), *Anglo-American Cataloguing Rules, 2nd ed.* (Ottawa: Canadian Library Association; London: Library Association Publishing; Chicago, Ill.: American Library Association, 1988).
- Guidelines for the Use of Field 856 (Network Development and MARC Standards Office, Washington, D.C.: Library of Congress, 1997)*.  
<http://lcweb.loc.gov/marc/856guide.html>
- B. Holt, 'Presentation of UNIMARC on the Web: new fields including the one for electronic resources', 64th IFLA General Conference, Amsterdam, Netherlands, 16-21 August 1998.  
<http://www.ifla.org/IV/ifla64/110-161e.htm>
- ISBD(ER) *International Standard Bibliographic Description for Electronic Resources: revised from the ISBD(CF): International Standard Bibliographic Description for Computer Files (UBCIM publications, New Series, 17. Munich: Saur, 1997)*.
- ISO 639-1:1988, *Code for the representation of names of languages* (Geneva: International Organisation for Standardization, 1988).
- ISO 639-2:1998, *Codes for the representation of names of languages - Part 2: Alpha-3 code* (Geneva: International Organisation for Standardization, 1998).
- ISO 8601:1988, *Data elements and interchange formats - Information interchange - Representation of dates and times* (Geneva: International Organisation for Standardization, 1988).
- E. Jul, *InterCat year-end statistics (E-mail to OCLC Internet Cataloging project list INTERCAT@oclc.org, 4 January 1999)*.  
[INTERCAT@oclc.org](mailto:INTERCAT@oclc.org)
- D. M. Levy, 'Cataloguing in the digital order', *Digital Libraries '95: The Second Annual Conference on the Theory and Practice of Digital Libraries, Texas A & M University, Austin, Texas, USA, 11-13 June 1995*.  
<http://csdl.tamu.edu/DL95/papers/levy/levy.html>
- C. Lynch, 'Searching the Internet', *Scientific American* 276 (3) (March 1997), 52-56.
- M. Münnich, 'German authority work and control', *Authority Control in the 21st Century, Online Computer Library Center (OCLC), Dublin, Ohio, 31 March-1 April 1996*.  
<http://www.oclc.org/oclc/man/authconf/muennich.htm>
- N. B. Olson (ed.), *Cataloging Internet resources: a manual and practical guide, 2nd ed.* (Dublin, Ohio: OCLC Online Computer Library Center, 1997).  
<http://www.purl.org/oclc/cataloging-internet>
- A. Sandberg-Fox & J. D. Byrum, 'From ISBD(CF) to ISBD(ER): process, policy, and provisions', *Library Resources and Technical Services* 42 (2) (1998), 89-101.
- V. T. Sha, 'Cataloguing Internet resources: the library approach', *The Electronic Library* 13 (5) (1995), 467-476.
- D. Vazine-Goetz, 'OCLC investigates using classification tools to organize Internet data', in P. A.

Cochrane & E. H. Johnson (eds.), *Visualizing subject access for 21st century information sources* (Urbana-Champaign, Ill.: University of Illinois at Urbana Champaign, Graduate School of Library and Information Science, 1998), 93-105.

M. Wolf & C. Wicksteed, *Date and Time Formats* (submission to World Wide Web Consortium (W3C), 15 September 1997).  
<http://www.w3.org/TR/NOTE-datetime-970915>

S. Zillhardt & F. Bourdon, *AUTHOR project: final report* (Paris: Bibliothèque nationale de France, 5 June 1998).  
<http://www.bl.uk/information/author.pdf>

## Credits

---

Chapter author: [Michael Day](#)

With contributions from: Emma Place

## 2.5 Subject classification, browsing and searching

### In this chapter...

---

- classification schemes
- keywords and thesauri
- staff issues
- browsing and searching
- future developments - automated solutions

## Introduction

---

The use of classification schemes, keywords and thesauri are central features of the formal resource descriptions provided by your service. The appeal of information gateways is based not only on the guaranteed high quality of the selected resources, but also on the facilities for subject-based access to the collection. In particular, information gateways typically provide access for both searching and browsing. Browsing (through a directory-like structure) is usually based on subject classification schemes or, exceptionally, thesauri. There are many such classification schemes from which to choose. You will need to decide which scheme suits the purpose of your gateway and the requirements of your target user group.

## Issues for gateway managers

---

This chapter should help you answer the following questions:

- do I want to use a classification scheme?
  - What are the pros and cons?
  - Which schemes are available?
  - How do I decide which one is the most appropriate scheme for my service?
  - Is it better to design my own scheme instead of using an existing scheme?
  - Can I adapt or extend existing schemes?
- is it useful to adopt a keyword system as well as a classification scheme?
  - What are the pros and cons of using controlled and uncontrolled vocabularies
  - What are thesauri?
- will my users require both searching and browsing facilities?
  - Is there an existing classification scheme which might be the best basis for a browsing structure or could a thesaurus or keyword system be adapted for this purpose?
  - How do I create a browsing structure from a classification scheme?
- how will my choices affect interoperability issues?

- how will my choices affect multilingual issues?

## Classification schemes

### What is subject classification?

Libraries have long experience of classifying resources, mainly books. The purpose of classification is to make it easier for users to find and retrieve resources. Subject classification is a method of describing resources by their subject. Universal classification schemes designed for use by libraries were first developed in North America during the nineteenth century. The most famous (and most widely used) scheme is the Dewey Decimal Classification (DDC) system, which was first produced for a small college library in 1876.

Classification schemes differ from other subject indexing systems, such as subject headings and thesauri, by trying to create collections of related resources in a hierarchical structure. The use of notations or codes facilitates the creation of hierarchical subject trees. For example, using UDC we can create the following hierarchy (adapted from McIlwaine, 1995, p. 17):

5	Natural science
504	Environmental science
504.05	Adverse effects of human activity on the environment
504.054	Effect of harmful materials. Pollution
504.054(44)	The effect of pollution on the environment in France

By building a hierarchical structure, a classification scheme enables users to look for related items which might otherwise be missed. This facilitates browsing, both within a physical library or online.

One advantage of an on-line system is that you can assign more than one classification number to a resource, since they do not need to be put in numerical order on a shelf; they can be (virtually) kept in two places at once. An Internet service can easily offer several different classification 'views' of the same resources.

### Types of classification schemes

Classification schemes can be broadly divided into:

Type	Characteristics	Examples
Universal schemes	General (covering all subject areas) Designed for worldwide usage	DDC (Dewey Decimal Classification) UDC (Universal Decimal Classification)
National general schemes	General in subject coverage Usually designed for use in a single country/language community	BC (Nederlandse Basisclassificatie) - Dutch SAB (Sveriges Allmänna Biblioteksförening) - Swedish
Subject-specific schemes	Designed for use by a particular (national or international) subject community	NLM (National Library of Medicine) Ei (Engineering Information Classification Codes)
Home-grown schemes	Designed for use in one particular service	Yahoo!

All of these classification types are used to some extent on the Internet (Koch and Day, 1997). Universal schemes like DDC and UDC are used by many Internet services and are readily available in machine-readable form. Subject services, however, are more likely to use a subject-specific scheme.

### Advantages of using a classification scheme for organising Web resources

The use of classification schemes offers one way of providing improved access to Web resources. It is not enough to build a collection of resources on the Web of a specific standard or relevant to a

particular audience. It is also necessary to organise and present those resources in such a way that the user can retrieve all the relevant resources quickly and easily. There are many Web guides which present resources in some kind of listing, either alphabetical or divided into ad hoc, constructed subject categories. These lists can soon become long and cluttered.

Classification schemes have therefore begun to replace less sophisticated ways of listing resources. A site which uses a classification scheme to organise knowledge demonstrates several distinct advantages over sites which do not (Koch and Day, 1997):

#### *1. Ease of browsing*

Classified subject lists can easily be browsed in an online environment. Browsing is particularly helpful for inexperienced users or for users not familiar with a subject and its structure and terminology. In addition, the structure of the classification scheme can be displayed in different ways as a navigation aid. The classification notation does not even need to be displayed on the screen, so an inexperienced user can have the advantage of using a hierarchical scheme without the distraction of the notation itself.

#### *2. Narrowing searches and viewing related resources*

When queries are limited to individual parts of a collection (filtering), the number of false hits is reduced, i.e. precision is improved. Classification schemes are hierarchical and therefore can also be used to get an overview of resources covering broader or narrower topics as you move up or down the hierarchy. This offers users the opportunity to view related resources which may be relevant to their information needs.

#### *3. Providing context*

The use of a classification scheme gives context to the search terms used. For example, the problem of homonyms (words which have the same spelling but a different meaning) can be partly overcome, because the context of the broader subject area or discipline will in most cases unambiguously indicate their meaning.

#### *4. Partitioning and manipulating databases*

Large classified lists can be divided logically into smaller parts if required.

Using an established or standard classification scheme has further advantages:

#### *5. Potential to permit multilingual access to a collection*

Since classification schemes often use language-independent notations (numerical or alphanumeric), these notations can be linked to as many of the available translations of the classification terms as you need. This offers the possibility of searching for terms belonging to a particular notation in various languages, and it also allows for the creation of browsing sections in more than one language. Other languages can be added later with very little effort, and without the need to classify the resources again. DDC and UDC have a good multilingual capability as the codes they produce are entirely numerical and their schedules have been widely translated (into nearly as many as 30 different languages). A version of a scheme in an appropriate language will not always be available.

#### *6. Improved interoperability*

The use of an agreed classification scheme could enable improved browsing and subject searching across databases.

#### *7. Greater stability*

An established classification does not usually become obsolete. The larger schemes are undergoing continuous revision, although they are normally also formally published in numbered editions. Some classifications may have to be changed when a new edition of a scheme is published, but it is unlikely that every single resource will have to be reclassified.

#### *8. Greater familiarity*

Some classification schemes are well known by a large user group. Regular users of libraries will be familiar with at least part of one or more of the traditional library schemes. Members of a subject community are likely to be familiar with their (subject-specific) schemes as well. Indeed some classification schemes are available in machine-readable form. Internet services which use established classification schemes may therefore have an advantage over those which use a home-grown scheme or none.



- A list of Web-accessible classification systems and thesauri is maintained at: <http://www.ub2.lu.se/metadata/subject-help.html>

### **Disadvantages of using a classification scheme for organising Web resources**

However, classification schemes also have some disadvantages:

#### *1. Splitting up logical collections of material*

Classification schemes often split up collections of related material, although this can be partly overcome with good cross-references and by assigning multiple class numbers to one resource.

#### *2. The illogical subdivision of classes*

Some popular schemes do not always subdivide classes in a logical manner. This can make them difficult to use for browsing purposes.

#### *3. Delays in assimilating new areas of interest*

Classification schemes, since they are usually updated through formal processes by organised bodies, often have difficulty in reacting promptly to new areas of study and changing terminology.

### **The most appropriate classification scheme for your service**

There are many factors to consider before choosing the most appropriate classification scheme for your service. Comparing the different types of scheme is one useful approach.

#### *1. Creating your own scheme versus using an existing scheme*

When a new gateway is being developed, you may be tempted to invent a new classification scheme for it. Inventing a new scheme has some advantages, but may also create new problems.

Advantages of creating a new classification scheme:

- A customised scheme, adapted specifically to the content and user groups of the gateway, should be able to meet all of its specific requirements. This should allow for easier and more consistent browsing. For example, there should be no unnecessary parts of the structure which would end up being unused.
- Home-made schemes are flexible and easy to change and therefore should be able to absorb new areas of interest relatively easily.

Creating a new classification scheme also has disadvantages:

- It is time-consuming - and therefore expensive - and requires extensive specialist subject knowledge.
- Even when the time and specialist knowledge is available, it is relatively easy to overlook something in a home-made scheme. For example, a gateway may find it difficult to fit a new term or hierarchy into its own scheme which was not considered when it was created. In addition, subject classification is a very subjective activity and this can easily lead to a lack of consistency.
- Custom-made schemes are not familiar to users, as existing universal or subject-specific classification schemes may be.

- Probably the main disadvantage is the almost complete lack of interoperability with other services and databases when it comes to subject description for browsing and searching.

Choosing an existing classification scheme avoids having to deal with some of the above issues. The scheme has already been made and it does not require any additional time or money to develop it.



Use an existing classification scheme, unless there is absolutely no suitable or adaptable system available or only schemes which cover a small part of the subject area. In this case it might pay to develop something completely new or adapt existing schemes which are only partly useful.

## *2. Established library classification schemes versus schemes developed for Internet usage*

The established library classification schemes have developed over a long period of time, sometimes as long as 100 years. This means that their conception of the world can be outdated and this may be reflected in the structure. For example, all universal schemes have had to take account of the rapid growth in electronics and computing in the second half of the twentieth century. Updating classification schemes takes a long time and sometimes the updated versions lack consistency, with new concepts being placed under illogical headings. Due to their size, the classification schemes do not update very often and, when they do, they tend to update one subject at a time. Traditional schemes can therefore be rather complex to use.

The good thing, however, about general library classification schemes is that they are universal schemes. They are built to classify an entire world with all its content. The schemes developed for Internet usage are of course relatively young, often developed over the last few years. This means that they are often still incomplete and continuously updating, trying to cover new subject areas as they go along. These schemes mirror the modern and changeable world. Sometimes they concentrate on a few areas of interest, ignoring the rest, sometimes they try to cover the whole world in the same way as the universal library classification schemes.

However, many home-grown schemes display severe weaknesses which hamper correct and efficient usage: failures in logic and hierarchy; incorrect subdivision of classes and application of multiple hierarchies; errors in terminology and in internal links and relationships between classes, and so on. There is also no requirement for subject services to use all layers of the classification hierarchy in an established system. Some current schemes organise material based on the first three levels only of a decimal scheme like DDC.

### **EXAMPLE**

Two good examples of classification schemes used for the Internet, the first an established one, the second home made:

- **BUBL LINK** is a comprehensive service covering academic resources in all subject areas. It uses the Dewey Decimal Classification (DDC) to classify documents
- **Yahoo!** is a commercial search service covering most popular subjects. Yahoo! uses its own universal classification scheme with 14 main categories

## *3. Universal classification schemes versus subject-specific schemes*

Universal classification schemes and subject-specific schemes are designed with different purposes in mind. A new gateway would need to choose a scheme relevant to the target audience for whom the service is being created. Where a gateway gives access to resources from all areas of knowledge, published throughout the world and in many languages, and intended to be offered to an international multi-disciplinary community of users, an existing universal scheme should be selected. If the service is a subject-specific one aimed at researchers within, say, the engineering community, it would be better to use a subject-specific classification scheme, if a suitable scheme is available. An alternative might be to use the appropriate part of a universal scheme.

Problems will occur for services covering subjects for which several different schemes exist (e.g. the earth sciences) or services which cover more than one subject area (e.g. the social sciences). In these cases, mapping and linking between schemes, the use of concordances for conversion, or extensions of a scheme may help.



### EXAMPLE

Two examples of subject-specific classification schemes:

- **SOSIG** (Social Science Information Gateway) uses part of UDC to generate a browsing structure (at the moment the categories are only displayed in alphabetical order)
- **EELS** is structured according to the subject classification scheme produced by Engineering Information Inc.

#### 4. National (monolingual) schemes versus international multilingual schemes

The choice between a national monolingual scheme and international multilingual schemes also depends on your subject and target group as well as on the purpose of the service. If a gateway only aims at a single user group within a country or at a specific language community and does not see any other potential users for the service, it could probably successfully use a national or language-based classification scheme. You would also possibly gain from the familiarity of a nationally-based scheme if you use one which is common in libraries. If, on the other hand, a gateway aims at a user group which is international (or which is intended to become international in the future), it would be better to use an international multilingual scheme, if one is available. If a gateway is thinking of cross-browsing or cross-searching with other gateways, it needs to consider the possibility of mapping to other schemes at this stage.

Note that some national schemes are available in a multilingual version, for example, the Nederlandse Basisclassificatie, which is the national scheme designed for use within the Dutch national cataloguing system. This scheme is available in English and (an adapted) German version as well. The English scheme is used on the Web in DutchESS; the German one is used by some German libraries which have adopted the Dutch Pica library cataloguing system.

### EXAMPLE

Using a national monolingual scheme:

- **Link Larder** - is a Swedish catalogue for quality assessed Internet resources (especially aimed at children). It uses the Swedish SAB for all subjects. The scheme is widely used in public and school libraries

Using an international multilingual scheme:

- **GERHARD** - the German academic Web index classifies all documents using the UDC classification from ETH Zürich in three languages

#### **Making your choice: issues to consider**

Your decision about the classification scheme you are going to use should also entail exploring the following important issues:

##### 1. *The scope and coverage of your service, and its primary target audience*

The scope of the service, its subject, language and geographic coverage, and its target user population should be the most important consideration in the choice of classification scheme. If the service includes all subjects and is aimed at a wide audience of Internet users, a universal classification scheme would be a good choice. If, however, the collection focuses on a limited subject area and there is a suitable international subject-specific scheme available, this should be used; if your service is a national service, you may want to consider a national general scheme. If no comprehensive scheme covering the geographic area or subject is available, a classification structure will have to be created especially for the service, either from scratch or (preferably) by extending an existing scheme.

##### 2. *Maintenance issues*

The decision concerning which scheme to adopt may also be affected by the level of familiarity that your staff have with a specific scheme, as well as by the maintenance level provided by the owner

of the classification system. If the staff are not familiar with the chosen scheme, this could slow down the growth of the gateway in the initial period.

### 3. *Quality, status and availability of the scheme*

Questions to be asked regarding this issue are:

- how do the considered systems compare in quality and controlled revision?
- is the scheme you want to use available in machine-readable form?
- is it available in the language you wish to use?
- is the scheme you want to use freely available for use on the Internet or do you need to acquire a licence?

### 4. *Interoperability issues*

The important consideration here is whether there are any mappings available between the candidate schemes and other established subject-specific or universal schemes which can secure interoperability to other services, now or in the future.

## CROSS REFERENCE

### [Interoperability](#)

### 5. *Costs*

How do the costs of the different schemes and methods compare? This includes costs for information specialists, technicians and (if necessary) translators as well as for servers and software being used.

The initialization of a service will require more investment, because all the issues discussed here need to be investigated, and the system chosen will have to be set up. When the service is up and running the costs will be lower.

## **Amending and mapping classification schemes**

Implementing classification schemes may present you with a number of issues. You may wish to adapt, restrict or extend the scheme you have chosen. There are also a number of very good reasons why you may want to map between multiple schemes. This section briefly summarises these issues.

### **Adapting a classification scheme**

For classification schemes to be effective as browsing aids in subject gateways, they need in some cases to be reduced in complexity and/or reordered.

A detailed table of the changes made should be kept, so that the locally used variant can be adapted easily whenever the original scheme is updated. For instance, when the hierarchy is rearranged, a mapping to the equivalent placings in the original scheme should be kept.

There are several ways in which classification schemes can be adapted:

#### 1. *Omitting empty classes*

A very unequal distribution of resources throughout a classification scheme can be confusing for the user and frustrate the browsing process. Omitting empty classes may be necessary in order to create a user-friendly browsing structure. If there are only a few empty classes or branches, the best policy is to mark the classes as empty in your browsing structure and navigation area (as done in EELS). The system will still appear as a coherent and logical whole. If there are many empty areas, the display could hide the empty classes. Our advice, however, is to classify the individual resources in as much detail as possible in the chosen system, but to display them for the time being in the broader/parent category. This allows for a fully expanded display as soon as there are enough resources for a meaningful finer substructure, without requiring any reclassification efforts. In any case, all resources should be displayed in order to keep consistency between browsing and searching the service.

## 2. *Rearranging hierarchies*

It may be necessary to rearrange the hierarchy to make the browsing structure easier to use. Sometimes the hierarchy needs a more logical arrangement to help users to find their way through it. Sometimes an important 'branch' deep down in the tree structure needs to be lifted closer to the top of the hierarchy so that it can be found more easily. In the end, if there is a potential conflict between the purpose of the gateway and the purpose of the classification scheme, it is the classification scheme which needs to be rearranged. If you are planning to include cross-browsing facilities in your gateway, rearranging hierarchies should be avoided as it complicates interoperability with other systems.

## 3. *Renaming captions*

Renaming captions is another way of adapting a classification scheme. A classification scheme may use complicated technical terms which would be difficult for the target audience to understand in a gateway designed for schoolchildren. In these cases, renaming adds value and user-friendliness to the service (cf. DDC for children and DDC for end-users). The renaming should be done in a similar way throughout the service in order to keep the service consistent and the language level the same.

### *Extending a classification scheme*

Sometimes an existing classification scheme is not detailed enough in particular areas or omits subject categories closely related to the gateway's coverage. If these are important areas for the gateway, then the classification scheme needs to be extended.

There are several different ways of extending a scheme:

- add a topical substructure to certain classes, without changing the existing classes; besides your own creations, bits and pieces from established more specific systems could be used
- add facets to the classification which allow subdivision of classes, e.g. a geographical or historical facet or a facet for document types or languages; the facets should preferably be taken from established systems
- 'glue' (parts of) an established system as a new branch on to your scheme to extend its topical coverage.

Again, document your extensions carefully so that you can identify these parts of your service and exclude them when carrying out operations based on your original scheme, such as adding resources from another service or cross-browsing. Remember, any mappings also need to be changed when changing your local scheme.

Consider that you have to maintain all the changes throughout the lifetime of your service. The extensions could be very useful and necessary for the service, but remember that they always involve extra costs, for instance in the form of extra work when adding resources to the service.

## **Conversion and mapping between classification schemes**

Mapping between different classification systems will become an increasingly important activity for subject services, in order to perform the following tasks (among others):

1. Conversion between different systems to incorporate records into a local structure or exchange of metadata, including automatically converting from existing classifications of documents (such as OPAC records, database records, documents in Internet services) into another scheme used in a subject gateway. An example is the mapping between DDC and UDC within the subject domains of economics and business for the SOSIG and Biz/ed projects (Hiom, 1998).
2. Support for the translation of categories and terms into other languages, to represent the different coverage of terms in different languages and to make up for the occasional lack of equivalent terms. A combination of translation and mapping may be the best way to accomplish multilingual vocabulary access and support. The EU Language engineering projects Acquarelle and Term-IT have been working in this area.
3. Extension of the classification structure by 'gluing' different systems into each other. This will be tested by the DESIRE II project together with OCLC. In 1995, a study was published exploring a mapping between DDC and the Mathematical Subject Classification MSC. (Iyer and Giguere, 1995)
4. Provision of cross-browsing between different services (which keep their classification

- systems unchanged).
5. Securing wide and future-proof interoperability options with different and maybe as yet unknown services.

### CROSS REFERENCE

#### [Multilingual issues, Co-operation between gateways](#)

Producing such a mapping is often difficult and time-consuming because of theoretical, conceptual, cultural and practical differences between the systems. Mappings have to apply many different types of equivalence; one-to-one relationships are certainly not sufficient. The mapping can be carried out between two or more systems or as a mapping to a universal system like DDC as a 'switching system' or 'interlingua'. The latter alternative is needed when trying to secure wide interoperability or when there is a small overlap between the used classifications.

If there are no 'official' conversion tables available, an improvement in the task of classification could still be made by extracting, from existing databases, linkages between different classification schemes or between indexing terms and classification for the same object, and using these linkages to construct a conversion algorithm.

In this field, neither theory nor practice is very mature. We recommend that you should seek advice and assistance from experts in the area.

### TIPS

- avoid as far as possible inventing anything on your own. This will help to ensure sustainability
- guarantee that there is a mapping of the scheme you are using to at least one important established classification system, whether international, subject-specific or universal. This makes your browsing structure interoperable and future-proof

## Keywords and thesauri

---

### Why use keywords?

In addition to the use of classification within an information gateway, information retrieval can be enhanced through the insertion of terms, or *keywords*, in a keyword field within each record. Such a practice has been common in the library world for many years as a means of aiding users to search abstracting and indexing services and library catalogues.

While classification of the records in an information gateway allows the presentation of groups of related documents in well-defined subject areas, keywords are used to give a detailed description of the concepts covered by the individual document and are mainly used as an aid to searching. The concepts covered by keywords are usually more specific than those of classes within a classification scheme, and consequently several keywords may be needed fully to describe a document. Individual keywords may therefore describe sub-topics within the page or site catalogued, whereas usually only one or two class numbers will be assigned to describe the overall subject content.

As noted elsewhere, keywords are generally applied to records as an aid to searching the catalogue (although they may also occasionally be used as a method of browsing - see the section on thesauri). Depending on the *type* of keyword system used and the policy adopted by the gateway in applying it, the added terms should improve the accessibility of individual records. They may also aid searchers by providing a feel for the philosophy and likely coverage of the gateway. An important function is to suggest to users new or more focussed terms with which they can search.

### Controlled versus uncontrolled

It is strongly recommended that some sort of keyword system be used when cataloguing sites for an information gateway, but it is important to decide whether or not to use a *controlled vocabulary* as the source of the keywords used.

A policy involving the use of *uncontrolled* vocabularies would consist of inserting into a keyword field terms relating to the subject content of the page or site which may or may not be contained

within the title of the document or included in any description that may have been applied to it. The keywords used will usually be suggested by an inspection of the site being catalogued or from the cataloguer's knowledge of the subject area. If the keyword field is included in your search, then such keywords should improve the recall.

The drawback with the use of uncontrolled keywords is that there are no standard, agreed terms for particular topics. This can cause problems not only with different spellings but with the use of different synonyms or near-synonyms to represent the same topic. Thus a search for the term 'labour relations' will not pick up records indexed with the term 'industrial relations'. Recall can be further improved by the correct and comprehensive application of a controlled vocabulary of standardised keywords.

As with classification systems, controlled vocabularies may be general in nature, such as the Library of Congress Subject headings (LCSH), or else be devised for one particular subject domain, such as the MESH vocabulary devised for the field of medicine. Since the majority of controlled vocabularies have been created for use with journal abstracting services, a suitable subject-specific system can usually be found by studying the major services in your subject area. Permission from the authors of the vocabulary should of course be obtained before using it within your gateway.

A problem with the use of controlled vocabularies is the constantly evolving nature of human knowledge resulting in the continual development of new terminology. As with classification schemes, major vocabularies periodically appear in new editions incorporating new terms, but it may happen quite frequently that a term cannot be found to describe the required content. There may also be problems with the degree of specificity of the scheme; that is, a term which is sufficiently specific may not be found.

The above problems can be alleviated by adding uncontrolled terms to records where a suitable controlled term cannot be found.

A consequence of using a controlled vocabulary is the need to make users aware of the vocabulary so that they are able to search on the allowed or preferred terms. This adds an extra complication to the gateway's interface, since the user will need to be able to search a version of the vocabulary for a suitable term if they are to make fullest use of controlled vocabulary indexing.

## CROSS REFERENCE

### [User interface implementation](#)

If the user is expected to search a copy of the vocabulary to select terms for a search, it is best to maintain a local copy of it which features only those terms which are present in your catalogue. This is particularly the case when the vocabulary is a large one and many terms within it would result in 'no hits'.

### **Indexing policy**

The search system your service uses and the search options you make available to the end users will, of course, have a critical effect on the users' experience of the service. However, as mentioned previously, the indexing policy of the gateway and how the keywords are added will also have a significant effect. As well as deciding whether to supplement terms from a controlled vocabulary with uncontrolled terms, an indexing policy should stipulate to what degree of specificity documents are to be indexed. The main issue here is that in cases where only keywords representing the main topics of the document are applied, the *precision* of a search can be increased if the search system has a mechanism for restricting searches to the keyword field.

It is generally recommended that you include all relevant keywords, including those occurring in the document's title and description, in the keywords field. However, if you decide not to restrict searches to a keyword field, you should be aware of the potential problems this might cause. Search results are sometimes displayed using ranking mechanisms which look at the number of times a searched-for keyword occurs in each record found and use this to order the results. Repeating terms already used within the description, for instance, may skew this process.

### **Thesauri - hierarchical controlled vocabularies**

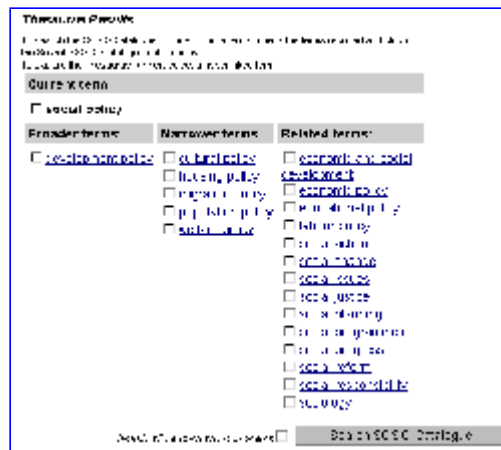
Controlled vocabularies may consist of large numbers of terms; they are also likely to comprise terms which are related to each other in various ways, particularly in broader/narrower relationships. Most of the major controlled vocabularies consequently have their terms arranged into hierarchies very similar to those of classification schemes.

**CROSS REFERENCE**

User interface implementation

The most common relationships between terms are:

- broader term (parent)
- narrower term (child)
- top term (the top of a branch of the hierarchy)
- related term (related but not broader or narrower)



The HASSET thesaurus produced by the Data Archive at the University of Essex, as used in the Social Science Information Gateway (<http://www.sosig.ac.uk/roads/cgi/thesaurus.pl>)

A hierarchical vocabulary or thesaurus makes it much easier both for the indexer to add relevant terms to the record and for the catalogue user to search on them. In principle, the user can begin at a top level term and browse down through the thesaurus until they come to a term closest to the topic in which they are interested. Some method for searching the thesaurus by keyword will also be available. In practice, a combination of searching the thesaurus and then browsing a small part will often give the user the best results.

The hierarchical structure is also useful in providing an overview of the structure of the subject domain (in a subject-specific system) for users who are unfamiliar with it, as with the browse structure derived from a classification scheme. It may also be possible to use a thesaurus in place of a classification scheme for browsing a catalogue, but the structure may not be as suitable for browsing as that of a classification scheme built for the purpose.



The figure above shows the medical gateway OMNI (<http://www.omni.ac.uk/search/thesaurus/>), which uses the MESH subject-headings to index its records. Selecting a particular term within the thesaurus produces a display of all records which contain this term.

Multilinguality

You may wish to create your own multilingual database which will allow users to perform searches within the catalogue, even though the original language of the record is unknown to them. Another approach would be to allow several separate databases in different languages to use the same thesaurus. As with classification schemes, it is possible that terms within a thesaurus can be represented by a unique identifier. If such a notation is used within catalogue records as well as or in place of the terms themselves, the display of keywords in records (or within the thesaurus) can be done in any number of different languages. However, any multilingual approach will require a great deal of time and effort - which is one reason why there are very few such multilingual services available.

### CROSS REFERENCE

#### [Multi-lingual issues](#)

### Staff issues

---

Subject classification and indexing are activities that in the library environment have been carried out by various trained professionals: subject specialists, cataloguers, information specialists or maintainers of (specialist) bibliographic databases. The quality of any browsing structure depends on the accuracy of the classification. The correct assignment of classification codes, keywords or thesaurus terms requires knowledge of the subject area as well as of the keyword system or classification scheme that is used. The process of assigning terms can be time-consuming.

Once you have decided that you want to add keywords and/or classification codes to the resource descriptions in your gateway, you will have to decide who among your staff has the necessary skills. This should be considered in relation to the question of who is going to be responsible for selection and/or cataloguing of the resources. One possibility is to let the same people select, index and catalogue the resources, which may be efficient; another option is to let people with different backgrounds and skills do the various tasks, which may make better use of the individual skills of various professionals.

A few possibilities:

1. Subject specialists, who select the resources, will usually have the required skills and/or experience with keywords and classification schemes, at least in their own subject areas.
2. Skilled (formal) cataloguers in some organisations will also be responsible for subject indexing. In other organisations their work will be restricted to the formal aspects of cataloguing, while index terms and so on will be added by a subject specialist. Whether cataloguers will be able to catalogue Internet resources, including subject indexing and classification, will depend on the situation in the organisation providing the service.
3. Trained librarians and information specialists, with various specific tasks within an organisation, will often have some skills in this area.
4. Another option is automatic assignment of classification codes or index terms. At the moment it is not possible to get the same high-quality results with automatic classification, without any intellectual human involvement.

### Browsing and searching

---

The methods for classification and subject indexing discussed so far should be evaluated in terms of their use in enhancing search and browse facilities in your gateway.

#### Browsing

Most services offer some kind of browsing facility. This may be an established classification scheme, a home-grown scheme, or some controlled vocabulary. This structure is typically presented to the user as a hierarchy starting from a list of terms, narrowing down till the user arrives at a list of resources. A list of resources may also be presented at each stage of the hierarchy.

Probably the best way to create a browsing structure is to use a classification scheme. Apart from providing a basis for the browsing structure, the numerical codes as well as the terms in whatever language they are available may be used for searching purposes as well. Numerical codes used for classification need not be displayed on the browsing pages. As noted previously, thesauri with explicit and complete hierarchical structures are also suitable for this purpose.

### CROSS REFERENCE

#### [User interface implementation](#)

## Searching

Many services offer 'advanced' search options, where searches on formal attributes (author, title) can be combined with terms specifying the subject of the resource. The latter may be uncontrolled keywords or terms taken from thesauri, subject headings, authority files and other vocabularies. Searching free-text descriptions may also provide an additional way of finding resources, either in combination with controlled keywords and/or classification codes, or in searches restricted to this field.

Classification schemes, although mostly used to provide a browsing structure, may also be used to enhance searching. These search options can be integrated in various ways in the user interface of your service. Sections of the classification scheme can be offered as a filter on the search, limiting the results of the query to a certain subject category of the database. The best way to do this is probably to offer a list of all alternative sections/classifications for selection, allowing the user to choose either one or several sections. An expert alternative would be to offer the classification field for direct searching with a truncation option, if the notation is made visible. On the browsing pages a search option could be offered limiting the search to the currently viewed class and the subclasses below. EELS and Yahoo! are examples of this approach.

Harvesting the documents in your service (and/or in your subject area in general) and providing a full-text index are other ways of expanding the services offered by your gateway. The user could choose to search either the record descriptions and/or the full text database. The latter would of course increase recall (even dramatically), but reduce precision. One example of cross searching a catalogue with a harvested index can be seen at <http://eels.lub.lu.se/aeels/search.html>

### CROSS REFERENCE

[Harvesting, indexing and automated metadata collection](#)

## Cross-browsing and cross-searching

Some subject areas are currently covered by more than one gateway; for example, engineering is covered by both EELS, EEVL and AVEL. This can be confusing for the users, who will have to have extensive knowledge about all existing gateways, to be able to decide which one(s) are most likely to answer their question. It is possible that one gateway may be more suitable for one subtype of resources than another, but users will have to compare various gateways, to get to know their strong and weak points, their exact coverage, biases and so on. The same problems arise for people interested in inter-disciplinary resource discovery. A possible way out of this dilemma from the service's point of view is to opt for more co-operation with other services in the same subject area. One way to co-operate is to enable the cross-searching and/or cross-browsing of gateways.

### CROSS REFERENCE

[Co-operation between gateways](#)

Cross-browsing two or more gateways is potentially a useful way of combining logically separate or distributed services, but it is difficult to achieve in practice. The gateways have to use identical classification schemes and the classification codes must be the same, so that a combined service can be generated, enabling a user to browse everything within the same virtual space; if identical schemes are not used, this becomes extremely difficult, if not impossible. Furthermore, classification is often a subjective activity and this would affect how combined subject gateways could be browsed. Nevertheless, cross-browsing through visible links between the browse sections of two or more gateways, without hiding their independence, can be accomplished by mapping methods as described previously; DESIRE II is currently testing different methods.

Cross-searching is relatively easy to provide in a networked environment, especially where the same search and retrieval protocols are in use. The resource description format has to be similar, though, and fielded search requires in addition semantic equivalence between the content of the fields in all services. Cross-searching has been tested by the ROADS project and can already be implemented in gateways based on the ROADS software (Kirremuir et al., 1998).

Cross-searching of information gateways poses a problem for the use of controlled vocabularies. As with cross-browsing using classification schemes, cross-searching only becomes possible if either the different catalogues use the same controlled vocabulary or if a mapping has been made between two or more different schemes. The latter possibility poses the same problems as are found when cross-mapping classification schemes and clearly it would be easiest if agreement



could be reached on the best vocabularies to use within particular subject areas.

Cross-searching and cross-browsing are more extensively covered in the Interoperability chapter. The User Interface Implementation chapter will tell you more about how to present browse and search facilities in your user interface.

### CROSS REFERENCE

[Interoperability](#), [User interface implementation](#)

## Future developments - automated solutions

---

### Automatic classification

As traditional classification is a time-consuming and expensive process, it is obvious that investigations into the use of automated solutions are worthwhile. At the same time, classification is an activity where a significant level of human expertise, abstract thinking and understanding is needed and this is not easy to replace by artificial intelligence or expert systems. There are no known examples of traditional library classification being undertaken completely by computer software. Knowledge structuring on the Internet has to cope with far larger numbers of resources, exponential growth rates and a high risk of changes occurring in documents which already exist.

This is the background to a growing number of research projects and experimental systems which are trying to support knowledge-structuring activities on the Internet with automatic methods. Most of these projects use methods of derived indexing, i.e. they extract information from the documents and then use it for structuring tasks.

Automated classification will probably not replace intellectual classification as far as quality subject services are concerned, but will rather support and complement selection and subject indexing efforts. Intellectual classification is always needed to validate and improve the automatic methods. However, robot-generated databases, as an add-on to quality services in a subject area, will be automatically classified. One practical goal in DESIRE II is to explore simple applications of automated classification methods on a robot-generated subject index to the Web. Many different tests will be carried out on the 'All' Engineering (AE) robot-generated database of engineering documents from the Internet. The effort required will be studied and the resulting outcomes evaluated. A pilot service of the 'All' Engineering Web index will offer a full classification and browsing structure with the most suitable solution found during the project. In addition, a comprehensive state-of-the-art report on projects, methods, alternatives and problems concerning automatic classification will also be presented. The results of DESIRE II will be included in the next edition of this handbook.

### Clustering

Clustering is a method which, like classification, aims to bring together groups of closely related documents. However, clustering is an automatic process, which groups documents according to specific criteria expressed in an algorithm. The groups are normally not (hierarchically) related to each other and are of very different sizes. The subject covered by a cluster is very hard to describe. Every time that new documents are added to the collection the clusters have to be calculated again and the outcome can be different. Documents can frequently move to other clusters. Clustering methods (which is a form of derived, a posteriori classification) should rather be compared with methods of automatic classification using established (a priori) classification systems used to assign classification to documents. Clustering is not suitable for presenting a stable structure for browsing large gateways in which documents need to be grouped into clearly defined and related subject sections; indeed, it is not meant to be used for that purpose.

## Further information

---

A more detailed analysis of the use of classification schemes in Internet resource description and discovery and a list of services using them can be found in the DESIRE I report produced by Koch and Day (Koch and Day, 1997). This report describes the use of several classification schemes on the Internet in some detail and provides an introduction to the use of automated classification techniques on the Internet.

Another useful Web page which lists some Internet-based services that use classification schemes for organising resource discovery services is Gerry McKiernan's Beyond Bookmarks page (McKiernan, 1996 and ongoing).

## Glossary

---

- Assigned indexing** Manual addition of meaningful terms to the records in a gateway to facilitate searching, usually taken from a pre-existing controlled vocabulary (see also derived indexing)
- BC** Nederlandse Basisclassificatie (Dutch Basic Classification, a Dutch national classification scheme used in the Pica Shared Cataloguing System.)
- Browsing** Information retrieval by navigating through a set of Web pages containing lists of resources grouped by subject
- Cross-browsing** Browsing, where the Web pages contain resources from more than one gateway
- Cross-searching** Searching, where the search takes place across more than one gateway
- DDC** Dewey Decimal Classification
- Derived indexing** Automatically extracting a list of terms from the documents in a collection to facilitate searching (see also assigned indexing)
- EELS** Engineering Electronic Library, Sweden
- Ei** Engineering Information
- Free-text searching** Searching using uncontrolled vocabulary, such as that found in titles, abstracts, or full text.
- LCC** Library of Congress Classification
- LCSH** Library of Congress Subject Headings
- MeSH** Medical Subject Headings
- NLM** National Library of Medicine
- OPAC** Online Public Access Catalogue
- Precision** The number of relevant documents retrieved divided by the total number of documents retrieved.
- Recall** The number of relevant documents retrieved divided by the total number of relevant documents in the collection.
- SAB** Sveriges Allmänna Biblioteksörening
- Searching** Information retrieval by entering one or more keywords into a search engine
- Thesaurus** A device for vocabulary control, usually for a specific subject area, indicating preferred terms, non-preferred terms, and semantic relations between terms; the terms are in ordinary human language.
- UDC** Universal Decimal Classification

## References

---

- Biz/ed, <http://www.bized.ac.uk/>
- DESIRE, <http://www.desire.org/>
- EELS, <http://eels.lub.lu.se/>
- OMNI, <http://www.omni.ac.uk/>
- SOSIG, <http://www.sosig.ac.uk/>
- T. Koch, *Controlled vocabularies, thesauri and classification systems available in the WWW. DC Subject*, <http://www.ub2.lu.se/metadata/subject-help.html>
- D. Hiom, *Mapping classification schemes (Bristol: SOSIG, 1998)* <http://www.sosig.ac.uk/desire/class/mapping.html>
- E. Miller, P. Miller & D. Brickley, *Guidance on expressing the Dublin Core within the Resource Description Framework (RDF)*, 1999 <http://www.ukoln.ac.uk/metadata/resources/dc/datamodel/WD-dc-rdf/WD-dc-rdf-19990427.html>
- J. Kirriemuir, D. Brickley, S. Welsh, J. Knight & M. Hamilton, 'Cross-Searching Subject Gateways - The Query Routing and Forward Knowledge Approach', *D-Lib Magazine (January 1998)*. <http://www.dlib.org/dlib/january98/01kirriemuir.html>
- T. Koch & M. Day, *The role of classification schemes in Internet resource description and discovery (DESIRE project: UKOLN, Bath, 1997)*. <http://www.ukoln.ac.uk/metadata/desire/classification/>
- T. Koch, 'Nutzung von Klassifikationssystemen zur verbesserten Beschreibung, Organisation und Suche von Internet Ressourcen', *Buch und Bibliothek 50:5 (1998)*, 326-335. <http://www.ub2.lu.se/tk/publ/bubmanus.html>

T. Koch, A. Ardö & L. Noodén, 'The construction of a robot-generated subject index', *EU Project DESIRE II D3.6a, Working Paper 1*, 1999.  
<http://www.lub.lu.se/desire/DESIRE36a-WP1.html>

T. Koch & D. Vizine-Goetz, 'Automatic Classification and Content Navigation Support for Web Services. DESIRE II co-operates with OCLC' in *Annual Review of OCLC Research 1998 (1999)*.  
[http://www.oclc.org/oclc/research/publications/review98/koch\\_vizine-goetz/automatic.htm](http://www.oclc.org/oclc/research/publications/review98/koch_vizine-goetz/automatic.htm)

T. Koch, *Controlled vocabularies, thesauri and classification systems available in the WWW (ongoing)*.  
<http://www.ub2.lu.se/metadata/subject-help.html>

I. C. McIlwaine, *Guide to the use of UDC: an introductory guide to the use and application of the Universal Decimal Classification*, rev. ed. (The Hague: International Federation for Information and Documentation (FID), 1995).

G. McKiernan, *Beyond bookmarks: schemes for organising the Web* (Iowa State University, 1996 and ongoing).  
<http://www.iastate.edu/~CYBERSTACKS/CTW.htm>

---

## Credits

Chapter authors: [Phil Cross](#), [Michael Day](#), [Traugott Koch](#), [Marianne Peereboom](#), [Ann-Sofie Zettergren](#)

## 2.6. Collection management

---

### In this chapter...

- the importance of keeping collections up to date
- methods for maintaining collections
- what do those error codes really mean?
- a link checking case study: SOSIG
- creating a collection management policy
- priorities for administrators

---

### Introduction

This chapter will look at some of the day-to-day administrative tasks required for running and maintaining an information gateway and the staff effort required for these tasks.

Whilst setting up and configuring a database for a gateway is labour intensive, it is a one-off task. The longer-term and time-consuming work is involved in creating and maintaining the collection: notably, in keeping the records up to date and error free. An out-of-date collection of resource descriptions is little use to anyone and may even be potentially harmful to users. It is important that sufficient staff effort is allocated for regular housekeeping duties, the main ones being:

- checking that resources are still available and links within records are still correct
- making sure that descriptions of resources are up to date and still adequately reflect the content of the resources themselves

The Internet is a volatile and fast changing environment; resources and information that are available today may not be available tomorrow. It has been estimated that at any one time between 5 and 8% of the Web's content is unavailable (Pitkow, 1998). There may be a number of reasons for resources not being available, ranging from networks being out of action, servers being out of order, or information being updated, to the resource's being removed permanently from the network. Whatever the reason, resources that are not available should be removed from your collection (if only on a temporary basis while the problem is solved).

Similarly, Internet resources do not tend to be static; they grow and change on a regular basis. Unless resource descriptions are checked on a routine basis, you may find that the records bear no resemblance to the resource itself, which may have changed or expanded beyond recognition within a few months or weeks.

## Maintaining collections

---

There are various tasks involved in making sure that an information gateway's collection maintains its integrity:

- validating records (spell checking, etc.) to ensure that the record is accurate
- link checking records to ensure that resources are still physically available
- updating resource descriptions to ensure that the record still adequately reflects content of the resource or Web site

### Validating records

A basic housekeeping duty is to ensure that catalogue records are as accurate as possible, not only in terms of the factual information they provide about a resource, but also in terms of the content of the record itself, e.g. making sure they do not contain spelling mistakes, that cataloguing guidelines are adhered to, etc. There are various internal procedures which can help gateways maintain accuracy within their records. These include:

1. Spell checking records. This can be done manually; some gateways employ staff to check and edit records before they are added to a live database. A less time-consuming way would be to use an automatic spell checker; however, there can be problems with spell checkers understanding discipline-specific or technical terms.
2. Cutting and pasting URLs and other pieces of factual information to avoid the possibility of typing errors.
3. Authority files. The use of lists of controlled terms and vocabulary can help enormously to cut down spelling mistakes and ensure consistency within the records.

For further information on ensuring accuracy and consistency within the collection see the chapter on cataloguing.

### CROSS REFERENCE

[Cataloguing](#)

### Link checking

Much of the information available over the Web is intentionally ephemeral in nature, designed only to be useful in the short term (e.g. TV listings, news bulletins, price lists). The average life span of a Web document is estimated at around 50 days, with HTML files being modified or deleted more frequently than images or other media (Pitkow, 1998). Gateways generally try to ensure that the resources they catalogue will have a degree of longevity and often include URL stability as one of their selection criteria. However, the inconstant nature of the Web means that it is still necessary to check resources regularly and update the records of those that have moved, are temporarily unavailable, or have been permanently deleted from the Internet. It is important to have collected contact information about the administrators or maintainers of the sites on which the resources reside. When a resource is unavailable, sending an email message to the administrator is often the quickest way to find out what the problem really is and whether or not it is temporary or permanent.

Automatic link checking software is available to help gateways keep a check on the resources described within their catalogues. The programs generally work by checking each of the URLs (often by requesting the 'HEAD' files of the pages) and compiling a report of any errors they find. The software can normally be scheduled to run at regular intervals (ideally at least once a week) and can be set to run at 'quiet' times, e.g. overnight, to reduce the load on the network. Once the error report has been generated, it usually then requires human effort to go through the report and decide which of the resources should be edited or removed from the catalogue. Working through an error report is much like detective work; you need to use patience, information finding skills and knowledge of the Internet to track down the problems and put them right.

As well as commercial software packages there are a number of link checking programs available in the public domain (freely available) or as shareware packages (for a small fee).

For a listing of some link checking shareware programs available see:

- [Link Checker Tools](#)

**What do those error codes really mean?**

You will sometimes see error codes when you are attempting to connect to Web pages or looking at the output of link checking reports. These are HTTP status codes and whilst they appear to be frustratingly cryptic they can tell you a lot about the type of problem that you are encountering.

**404 - Page Not Found**

This is the most common error code that gateway administrators will come across. Web site maintainers often change the structure of their sites, as the information they provide grows or as the maintainers get new ideas about how to arrange and present the information. One of the most common reasons for a 404 error is simply that the resource has been moved to a different part of the site. To find the new location you can often systematically move up the directory structure of the URL deleting the text before each trailing slash (/) until you find a link to the resource. Sometimes the resource may have moved to another Web site altogether (this often happens when the resource is located on a commercial site); it is worth doing a search on one of the big search engines (such as Alta Vista) to try to locate its new address. In the worst case, the resource has been deleted permanently and the record should be removed from the collection. If you cannot locate the resource simply by looking around the site, an email message to the administrator will often solve the mystery.

Some of the other frequent error codes are:

Error Code	Problem	Possible Reason and Action
401	Unauthorised Request Access	The resource may be protected by a username and password - contact the maintainers for more information.
402	Payment Required	The request requires a charge to be applied to the transaction.
403	Forbidden	Access to the directory is forbidden. The resource may no longer be available for public access or the Web site administrator may have changed the directory permissions by mistake!
500	Internal Error	These types of error messages are very frustrating, as it is often hard to pin down what the problem is. It may be a problem caused by attempted execution of a CGI script. The best course of action is to monitor it as a problem and email the maintainer of the site for more information about the nature of the problem and to find out whether it is temporary.
501	Not Implemented	The server does not support the method being requested.
503	Server Busy	The server is unable to process the request for the page because of the high number of other requests. These tend to be temporary errors; try again at another time.

**A link checking case study: SOSIG**

SOSIG uses the link checking software that is supplied as part of the ROADS system. The program is scheduled to run automatically just after midnight on Sunday when the network traffic is generally low. The program runs through each of the URLs in the SOSIG database (over 7,000) and for each it requests the HEAD file from the page. If the request is successful the software moves on to the next URL; if it encounters a problem it writes the URL and the unique ID number for the record into a file. Once the link checker has processed all of the URLs, the problem resources are sorted and presented according to the error codes discussed in the section above. The error report is made available through the SOSIG online administration centre (see Figure 1); additionally a copy is emailed to the SOSIG staff responsible for processing the report.





Figure 1 SOSIG Link Checking Summary Report

SOSIG currently has one member of staff assigned to link checking, who spends approximately one day a week going through the report and updating or deleting records as appropriate. As the number of records in the collection grows, so does the number of problem resources, and it is likely that the amount of time required to maintain the collection will increase over time.

The errors reported are given an order of priority and the '404 Page Not Found' problems are dealt with first of all. These are probably the most straightforward of the errors; either the resource has moved and the record has to be edited to have the new address or it is no longer available and it needs to be deleted from the database. Either way, having error pages appear when users try to connect to resources is likely to reduce their confidence in the collection.

The next errors dealt with would be any errors to do with authorisation (error 401), payment (error 402) or permissions (error 403). These errors are not as common as the 404 errors and they tend to appear when a resource that had previously been publicly available is now restricted to use within an organisation or community and some form of payment or authorisation is required. These problems may become more common as the Web matures and commercial practices become more established. Occasionally the problem is simply that the Web site administrator has inadvertently changed the permissions on the directory and is unaware that there is a problem. SOSIG has found that the best way to deal with these problems is to get in touch with the maintainers of the resource by email and ask what the situation is; generally replies return within a day and the record can be dealt with appropriately.

The final errors that are dealt with are the 500 errors, generated by the server from which you are requesting the resource. They tend to be more unpredictable and it is usually quite difficult to pinpoint the problem; often URLs listed as giving 500 errors are working perfectly well when checked again. The reason for this may be because that the server was undergoing maintenance or updating when the link checker requested the URL. SOSIG tends to monitor 500 errors over a few weeks and an email message will be sent to the maintainers of those resources that persistently record an error. The ROADS link checking software does have a feature which allows you to automatically delete URLs that are consistently unavailable, but this is not used as it is felt that the 500 errors are too unpredictable and staff prefer to make a judgement on each resource.

For more details of the link checking software and the ROADS software in general see:

- [ROADS Project Software/Documentation Server](#)

### Updating resource descriptions

The dynamic nature of the Web is a problem when it comes keeping manually catalogued records of resources up to date and relevant. Web documents, unlike their printed equivalents, are very easy to edit and modify; studies have shown that most Web pages are not static but expand and evolve over time. For a gateway's collection to maintain its integrity and usefulness, the records must also reflect the changes in the resources. This is a time-consuming job that requires ongoing staff effort to be assigned to the task.

There are a number of steps which gateways can take to help to identify and review resources that need their descriptions to be updated:

1. Making full use of administrative metadata such as review-by dates. When records are created, a date can be added by which this record should be reviewed. A simple script can pull out all of the records that require reviewing at any particular time.
2. Using automated processes to email resource maintainers to ask whether there have been any changes to the resource since the record was created.
3. Using automated processes to delete time-dependent resources, e.g. conference announcements.
4. Using Web page tracking tools (such as Mind-it <http://mindit.netmind.com/>) to monitor changes in resources (these generally report changes when the size of the file is altered).
5. Taking the opportunity to update descriptions of records that are being edited as a result of running a link checker.

## Creating a collection management policy

---

The Web has often been described as a 'moving target'; it is constantly changing and expanding and trying to catalogue its content is a difficult business. Gateways need to think about what they are trying to provide for their users: a catalogue of the entire Web or a focused collection of selected material? A previous chapter on quality selection criteria has dealt with the need for gateways to consider formalising a Scope Policy to help clarify the type of service they are offering. It will also be helpful to think about a policy for managing collections. A collection management policy will allow you to formalise not only the scope and selection criteria for a gateway but also deselection criteria, that is the principles under which you may choose to edit or delete records from the collection. A collection management policy might include:

### Guidelines for deselecting a resource:

- if the resource is no longer available
- if the currency or reliability of the resource has lessened
- if another Internet site or resource offers more comprehensive coverage

### Guidelines for editing a record:

- if the information content of the resource has changed so that the resource description and keywords need to be updated
- if any of the factual details of the resource have changed (e.g. new admin email, new short title)
- to correct any errors made in the original record

Collection management policies may change over time to reflect the changing nature and content of the Web. As more resources become available it may be necessary to delete entries from the collection, replacing them with more suitable material.

For examples of gateway collection management policies see:

- [ADAM Collections Policy](#)
- [SOSIG Collection Management Policy](#)

## Priorities for administrators

---

When one is faced with limited time and resources, there will always be a conflict between building up the gateway collection and adequately maintaining the existing collection. In order to continue to offer useful services, gateway administrators need to ensure that they balance effort spent in creating new records with preserving the integrity of the current collection. It is advised that gateways make as much use as possible of automated tools to monitor and track changes in resources, so that any human effort is directed at the more intellectual tasks of revising and correcting records.

## Glossary

---

**ADAM** Art, Design, Architecture and Media gateway (UK)

**authority file** cataloguing tool that offers the cataloguer a set list of options from which they must choose to fill a particular field - ensures consistency of entry within catalogue fields

**ROADS** Resource Organisation And Discovery in Subject based services. eLib funded project developing software for use by Internet subject services.

## References

---

Mind-it by NetMind, <http://mindit.netmind.com/>

ROADS, <http://www.roads.lut.ac.uk/>

SOSIG, <http://www.sosig.ac.uk/>

W. Koehler, *'Digital Libraries and World Wide Web Sites and Page Persistence'*, *Information Research Volume 4 No. 4 (June 1999)*.

J. E. Pitkow, 'Summary of WWW Characterizations', in *Proceedings of the Seventh International World Wide Web Conference, 14-18 April 1998, Brisbane, Australia (Elsevier Science B.V., 1998)*.

## Credits

---

Chapter author: [Debra Hiom](#)

With contributions from: Phil Cross and Emma Place

## 2.7. Working with information providers

### In this chapter...

---

- identifying the key information providers for your gateway
- building and maintaining relationships with information providers
- involving information providers in the metadata creation process

### Introduction

---

One of the most time-consuming, and therefore costly, tasks for information gateways is maintaining up-to-date descriptions of relevant resources. Identifying and describing quality resources is critical for the gateway. One possible means of making this process more efficient is to involve the 'information providers' (otherwise described as 'publishers' or 'resource owners') in the metadata creation process and to encourage them to contribute to the content of the gateway. This benefits the gateway in terms of saving costs and at the same time helps ensure the currency of the information held by the gateway. The benefit to the information provider lies in improved dissemination of their information. This is an alternative approach to the creation of resource descriptions 'by hand', where metadata is created centrally by the information gateway's own staff, or by library staff who are working within other institutions, or by subject experts.

These various methods are in use to a greater or lesser extent in existing gateways. In the UK, for example, the Resource Discovery Network gateways have most of their metadata created by gateway staff or subject experts, but services such as the Arts and Humanities Data Service rely to a much greater extent on resource creators inputting data to the gateway.

In the case of those gateways where metadata is created automatically by harvesting or crawling the web, it is also possible to involve information providers; this may be by agreeing procedures for identifying relevant material automatically, or by the information provider's alerting the gateway to new or updated data.

In this chapter we will look at some of the issues which arise when gateways and information providers work more closely together. We will consider the benefits of this approach but also note any disadvantages.

### Identifying information providers

---

Whatever method of metadata creation is followed, a primary task for any gateway is to identify the key information providers in its field. These key providers may be individuals, groups or institutions who are creating or have some level of ownership of high quality resources. In the case of Higher Education funded gateways, the key information providers may be individual researchers, university departments, publishers, scholarly societies or commercial organisations working in the relevant subject area.

The key providers may vary considerably as regards:

- the volume of relevant resources they produce
- the rate at which resources are updated, i.e. volatility of resources
- whether they create metadata themselves at source for their own resources

Taking these factors into account, the gateway will need to consider the overall profile of its key information providers in relation to gateway policy for metadata creation. The gateway needs to consider its own policy by asking:



- what is the optimum number of records in the gateway? Is there an imperative need to build up the volume of records in the service?
- at what level of granularity are resources being described? Can information providers help the gateway to describe resources at a finer level of granularity?
- how rich is the metadata in the gateway? If the gateway wishes to produce rich metadata, then contributions from providers may need to be enhanced. Careful consideration needs to be given to the cost of enhancement as compared with creation from scratch.
- are there benefits in building relationships with providers over and above the value of the imported metadata? Key providers may be key users whom it is beneficial to have on board.

It will also be useful to look at the wider picture and consider the cost of involving information providers. In order to justify setting up complex systems, the gateway will want to be assured that information providers can contribute a significant quantity of metadata. It may be that, to create economies of scale, gateways will need to co-operate with one other in setting up common methods for importing metadata from information providers. It is also likely that the information providers themselves will be contributing to a range of gateways and they will want a common procedure to cover all gateways. Such procedures would need to be flexible enough to allow for differing practices among information providers while following internationally accepted standards and protocols which can be clearly defined.

### **Building relationships with information providers**

---

Having identified key providers and decided that they can contribute to the content of the gateway, the gateway can then build on this information in various ways.

#### **Monitor key information providers**

At the simplest level the gateway can ensure that a system is in place to monitor regularly the web sites of key players. This may involve guidelines for staff and varying degrees of automated monitoring. For example, staff may bookmark sites to check regularly or use a URL-minder to notify them of changes made to key sites.

#### **CROSS REFERENCE**

[Resource discovery](#)

#### **Enable submission of metadata**

The gateway can offer a means for information providers to provide data about new resources. This may be a 'Submit a Resource' form on the gateway Web site.

#### **EXAMPLE**

##### **Example of encouraging submission of metadata from information providers**

Within DutchESS, resources are selected by subject specialists in the participating libraries on the basis of quality and relevance to the academic community. On the Web site there is a page for 'adding a resource' which asks:

*Do you want to contribute a new resource to DutchESS? Use this form to let us know. Your suggestion will be submitted to one of our subject specialist[s]. If the resource is according to the scope policy and quality criteria of DutchESS it will be added to the database.*

- <http://www.konbib.nl/dutchess/index.html>

#### **Information providers create the metadata**

Gateways can offer metadata guidelines for providers who publish large numbers of relevant resources, so that they can create the metadata required. The metadata can then be automatically transferred to the gateway. Metadata may be manual, using a web based form, or semi-automated, using one of the available metadata creation tools. (CROSS REFERENCE metadata creation chapter)

**EXAMPLE****Examples of gateways using metadata created by trusted information providers**

A full-text electronic journal, SocRes Online, undertook an experiment with SOSIG, whereby the journal created metadata for each article, which was then automatically imported into SOSIG. Quality guidelines were agreed with the journal. This saved SOSIG staff considerable time, as they did not need to create records for the articles but simply needed to check the records that had been automatically created.

- <http://www.socresonline.org.uk/socresonline/>

Indoreg (Hansen and Hansen, 1997) is a Danish project looking at the bibliographic control of Danish Internet documents and is particularly concerned with the inclusion of Internet documents in the Danish national bibliography. The project concluded that 'self-registration' by authors or publishers would be needed if large amounts of information were to be registered. It recommended the use of Dublin Core for this self-registration and provided tools - a DC creator (based on the Nordic Metadata Project's DC creator) and a PURL server - that would facilitate this.

- <http://purl.dk/rapport/html.uk/>

**Endorsement by influential institutions**

It can be a condition of a grant that data resulting from funded projects should be deposited with a specified data repository. It might be that gateways could persuade funding agencies to insist that metadata is deposited with the relevant subject gateway.

**EXAMPLE****Example of institutionalised metadata creation**

It is a stipulation of the UK Arts and Humanities Research Board that funded projects deposit the data produced by the project with one of the service providers of the [Arts and Humanities Data Service](#) (AHDS).

This data may be in the form of a dataset or a catalogue record. The Archaeology Data Service, an AHDS service, recommends depositing a catalogue record if the data is dynamic, or if it is non-digital. As well as being a mandatory condition archaeology organisations, depositing data benefits the individual researcher. Benefits are summarised by the Archaeology Data Service under the following headings:

- professional recognition
  - avoiding duplication (of catalogue records in different locations)
  - building links between data sets
  - signposting data
- <http://ads.ahds.ac.uk/project/userinfo/deposit.html>
  - <http://www.pads.ahds.ac.uk/padsDepositorsGuide.html>

**Distributed collaborative cataloguing**

The future business model for metadata creation may lie with distributed collaborative cataloguing. This would involve an incremental approach to building up metadata for resources. The 'publisher' or 'owner' of the resource might create initial simple metadata, using the Dublin Core element set, for example. Services that wish to offer access to the resource might enhance this basic metadata, for instance with a description targeted at the ultimate users of the service. If the resource meets the criteria for description by the national library and inclusion in a national bibliography, then the national library might augment the records with subject headings and classification codes and align names and headings with the relevant authority files. Other interested parties might create unique identifiers (ISSN, DOI, etc.) or add metadata concerned with rights management or digital preservation. In this model the information provider becomes the first step in a chain of metadata

creators.

### CROSS REFERENCE

#### [Co-operation between gateways](#)

There are pilot projects investigating shared metadata creation where a 'workspace' is used to create metadata collaboratively. At present, these projects are looking at collaboration between specific partners in the metadata creation process, for example libraries working together or publishers working with national libraries and identification agencies. Within these projects metadata can be enhanced incrementally and imported or exported in a variety of formats.

### EXAMPLE

#### Examples of projects investigating shared metadata creation

##### [Biblink](#)

The BIBLINK demonstrator consists of the 'BIBLINK workspace' - a shared, virtual workspace for the exchange of metadata between publishers, National Bibliographic Agencies (typically national libraries) and other third parties such as the ISSN International Centre. The workspace will allow publishers to 'upload' metadata for electronic publications using email or the Web. National Bibliographic Agencies and third parties will be able to 'download' this metadata, enhance it in various ways and then 'upload' the enhanced metadata back to the workspace. The intention is that national libraries will use the enhanced metadata as the basis of a record in the national bibliography, if appropriate. Finally, publishers will be able to 'download' the enhanced metadata for use in their own systems. The metadata will be stored and exchanged in several syntaxes, including HTML, SGML, UNIMARC and the national MARC formats of the participating libraries.

##### [CORC](#)

CORC (Co-operative Online Resource Catalog) is an OCLC research project exploring the co-operative creation and sharing of metadata by libraries. CORC integrates recent metadata initiatives such as Dublin Core with MARC, enabling a more flexible approach to record creation. CORC emphasises the importance of exporting the records in syntaxes usable on the Web (e.g. HTML, XML/ RDF).

#### Community building

The gateway can build up a community of information providers. There may well be an overlap between providers and users of the gateway service, so this may be viewed as a marketing strategy. Traditional methods of dissemination (such as publishing, presentations, attending conferences) will form a basis for this activity. Growth of the community can be encouraged by invitational events for key players followed up by mailings and newsletters. A number of the eLib gateways in the UK have progressed from relatively simple catalogues of Internet resources to 'subject communities'. Depending on the business model by which the gateway is funded, membership of such a community of providers may confer benefits of preferential access costs or access credits.

### EXAMPLE

#### Examples of gateways establishing links with information providers and building communities

[EEVL](#), the engineering subject gateway, contains a range of information much wider than a search service; as well as a catalogue of selected 'quality' resources, it offers comprehensive searches of UK Engineering Web Sites, engineering e-journals and engineering newsgroups, and indexes to printed literature. As well as running the comprehensive Web site, EEVL organises training and awareness sessions.

[SOSIG](#) puts out calls to the social science community to request information regarding resources that they are publishing on the Internet. SOSIG now has good links with the academic social science community in the UK - as a result academics, government departments, the ESRC and others all send email to let SOSIG know when they put a new resource online. SOSIG has also run its [own conference](#) which brought together key information providers and users and established SOSIG at the centre of this community.

[Biz/ed](#) has responded to the most common information requests of their users by contacting key companies and organisations to request information. They have established links with organisations such as the Bank of England, the Office of National Statistics and Penn World Data. The gateway has created [primary resources](#) collaboratively with these organisations. Biz/ed has also contacted [companies](#) such as McDonalds, BMW and the Body Shop to ask for information to add to the gateway. See also: <http://www.bized.ac.uk/virtual/>

### Benefits and costs

---

There are a number of potential benefits resulting from information providers' providing metadata:

- cost saving
- assistance in keeping metadata up to date
- accuracy of details

These need to be balanced against:

- need to apply quality assurance
- effort spent supporting information providers
- instituting and maintaining processes for inputting data remotely

### Is this right for your gateway?

---

Some factors that may affect the emphasis the gateway gives to metadata supply by information providers:

- what is the likely scale of information provider contribution?
- how many individual resources will the information provider supply?
- what level of enhancement to metadata will be required to meet quality control criteria?
- is the service aiming at comprehensive coverage of an area?
- are information provider contributions seen as only as possible content for the gateway, or will information providers expect their data to be included (need to manage expectations)

### Conclusions

---

It is worth while building relationships with key information providers, especially as in many cases they are likely to be users of the information as well as contributors.

Gateways may judge that at present information providers cannot provide enough metadata to make it worth while setting up systems to import metadata. However, it seems likely that, as metadata standards mature, organisations owning resources will recognise the advantages of creating metadata for their own purposes which may be for administration, rights management, marketing, their own resource discovery systems or to pass along the retail chain. Gateways need to be ready to take advantage of changes in the pattern of metadata creation when (if) this happens.

Gateways will need to move towards a viable business model for metadata creation to ensure their longterm sustainability.

### Glossary

---

**AHDS** - Arts and Humanities Data Service

**CORC** - Co-operative Online Resource Catalog

**DOI** - Digital Object Identifier

**Dublin Core** - A metadata format defined on the basis of international consensus which has defined minimal information resource description, generally for use in a WWW environment.

**DutchESS** - Dutch Electronic Subject Service

**EEVL** - Edinburgh Engineering Virtual Library

**Elib** - The Electronic Libraries Programme (UK)

**ISSN** - International Standard Serial Number

**MARC** - MACHine Readable Cataloguing. A family of formats based on ISO 2709 for the exchange of bibliographic and other related information in machine readable form. For example, USMARC, UKMARC and UNIMARC.

**PURL** - Persistent Uniform Resource Locator.

**RDF** - Resource Description Framework

**SGML** - Standard general Mark-up Language

**SOSIG** - The Social Science Information Gateway

**XML** - Extensible Markup Language. A lightweight version of SGML designed for use on the Internet

---

## References

P. B. Hansen & J. Hansen, *INDOREG: INternet Document REGistration: project report (1997)*.  
<http://purl.dk/rapport/html.uk/>

---

## Credits

Chapter author: [Rachel Heery](#)

With contributions from: Emma Place

## 2.8. Publicity and promotion

---

### In this chapter...

- publicity and promotion - what are the issues?
- the power of well planned publicity
- traditional promotion and publicity activities
- online promotion and publicity
- combining promotion and publicity with other activities
- examples of effective publicity and promotion

---

### Introduction

Publicity and promotion are rarely at the forefront of people's minds when planning an information gateway, yet they are often essential ingredients for a gateway's success. Good publicity can help enormously to bring an information gateway to the attention of the people that really matter, i.e. the gateway's target users.

An effective publicity and promotional campaign takes time and effort to plan and deliver; it can also cost money. This chapter attempts to highlight some of the issues that should be considered when planning publicity and promotion activities.

---

### What are the issues?

The key issues at stake with publicity and promotion are: what is the intended audience?

- what kind of publicity and promotion is available?
- are all types of publicity worth while?
- how can a limited budget (time and/or money) be targeted most effectively?
- are there any failsafe methods for successful publicity and promotion?
- how can you retain the interest of your users?

---

### What is the intended audience?

You should think carefully about the audience which your publicity is intended to reach and win over. If you can characterise your user community carefully and target the publicity accordingly, it will be much more effective.

## What kind of publicity and promotion is available?

Publicity and promotional methods for gateways may be divided into three distinct forms: traditional media, electronic media activities and face-to-face activities. The underlying aims of each are very similar: to communicate to as many people as possible (ideally your target users) that your gateway exists and to convince them that they should use it. Once users find the gateway, then the quality of the resources should make them into repeat visitors.

### Traditional media activities

Traditional media activities are often overlooked as methods of publicity when Internet-related projects are planned. This is a shame, as they can be extremely powerful and far-reaching and can often produce the best results in terms of reaching the largest group of potential users. Traditional media can include paper-based materials (leaflets, posters, newsletters, papers, journals, magazines, etc.) as well as media such as television and radio.

### Paper-based materials

Paper-based materials fall into two distinct groups: publications in the form of journals, magazines and newspapers and paper publicity materials such as information sheets, leaflets and posters.

Publications can be used effectively to access concentrated groups of target users directly. If you place an advertisement in a specialist journal that is read by large numbers of your target users, the results can be well worth the money. Paying for publicity by means of advertising is not the only route (although it should be considered, as the results can be impressive, far-reaching and cost effective). Writing review articles in journals or newsletters can be a good way to get some 'free' publicity. Obviously, the time involved in writing such articles should be considered and costed. Nevertheless, articles written by gateway staff are often a very successful means of publicity.

### EXAMPLE

#### Example: Articles written by gateway staff

The following articles have been written by gateway staff and all act as good publicity materials, either directly or indirectly:

- Biz/ed:  
Catherine Sladen, 'Ethical Business', Business Review (April 1998). Catherine Sladen, 'Mergers and Take-overs', Business Education Today (May/June 1998).
- SOSIG:  
Debra Hiom, 'Around the table: Social scientists have their own favourite places on the Web', Ariadne 9 (May 1997). Debra Hiom, 'SOSIG: Providing access to internet information', Laser Link (Autumn 1998).
- OMNI:  
John Kirriemuir, 'A report on the third annual OMNI seminar: A cure for information overload', CTICM Update 8:2 (December 1997).

### EXAMPLE

#### Example: Advertisements placed by gateways

NMM Port:

To coincide with the launch of the Port information gateway, a number of advertisements were placed in maritime-related journals and publications. These included: Times Higher Education Supplement (16/04/99)

- Navy News (May 1999)
- Managing Information (April 99-6:3)
- History Today 49(5) (May 1999)
- Museums Journal (May 1999)
- Seabreezes 73 (641) (May 1999)

Another way for your gateway to appear in the user community literature is for it to be included or

referenced in other people's articles. Of course this may be harder to achieve as it requires people to know about and value the gateway. However, as a gateway matures and becomes a feature of the user community, this kind of publicity becomes more likely. Targeting known journalists or writers within your user community can also pay dividends and produce some favourable results. Consideration should be given to all contacts that people associated with the gateway may have.

#### E X A M P L E

##### Example: Articles written about gateways by non-gateway staff

- **Biz/ed:**  
*The Guardian newspaper (07/03/96)*  
 The Guardian's regular Web site review column contained a glowing review of the early Biz/ed information gateway.  
*Times Higher Education Supplement (02/04/99)*  
 An extensive review of one of the many features available from the Biz/ed information gateway: 'Website opens doors to No 11: Chris Johnston finds a site based on the economic model of the Treasury'. Although the article did not deal strictly with the information gateway resource catalogue, it did raise the awareness of the site as a whole. A good example of all publicity being good publicity!
  
- **NMM Port:**  
*The Times newspaper (11/05/99)*  
 The following article appeared in The Times newspaper: 'With no added salts: at last, an honest, unsentimental tribute to our maritime heritage' (by Libby Purves), containing several references to the Port gateway and its features.
  
- **Gateways in general (with reference to SOSIG):**  
*Times Higher Education Supplement (08/01/99)*  
 'Out of the morass: step through one of the Internet's subject gateways and you leave the information jungle behind, says Ayala Ocher'. An excellent review of information gateways in general and even references to the DESIRE project!

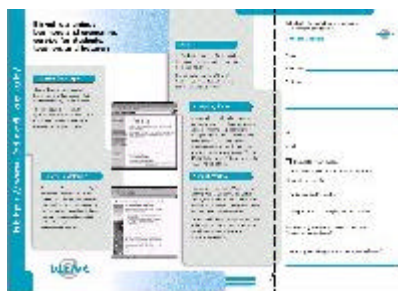
The benefit of carefully targeted articles or advertisements in your user community literature is that the materials immediately have context and are being viewed by people interested in the subject matter; this significantly increases the chances of people reading the article and subsequently visiting the gateway.

Other paper-based materials such as information sheets, leaflets and posters can also be very effective as promotional materials. Developing a visually attractive information sheet about your gateway and distributing it to key users can help to raise the profile of the gateway. Several gateways have used this idea to great effect. Promotional materials do not need to stop at information sheets. Bookmarks, mouse mats, mugs and T-shirts have all been used and have potential. Naturally, the exact kind of materials chosen may be largely dependent on cost and funding.

#### E X A M P L E

##### Example: Gateway information leaflets

Biz/ed: [PDF \(bized-flyer.pdf\)](#)



Port: [PDF \(port-leaflet.pdf\)](#)





OMNI: [PDF \(omni-leaflet4.pdf\)](#)



All of the materials above have been sent to key sections of the target user community (subject librarians, University libraries, subject-specific book shops and museums) who have been asked to display them where their users could see them. Having a Biz/ed information sheet available in the Social Science library near the networked computers has obvious benefits.

In several cases the promotional materials have been so popular that extra copies have been ordered by the people concerned.

Correctly targeting the recipients of promotional activities can produce a cascading effect, so that the targeted people then pass on their knowledge concerning the gateway to more people locally.

### Television and radio

Though perhaps not as appropriate for publicising gateways as some of the other media mentioned in this chapter, the use of television and radio does have enormous potential. Obviously the idea of placing a commercial for your gateway on the television or radio may be in the realms of science fiction, but getting the gateway mentioned as part of another programme may be a more down-to-earth ambition. This is especially true with the recent growth in popularity of Internet-focused programmes.

Gateways are more likely to get mentioned if they are well established, by coming to the attention of television and radio programme producers and researchers. Well placed contacts can also help to raise the profile of a gateway within the relevant circles.

### Electronic media activities

#### Search engines and directory listings

It goes without saying that an information gateway should make sure that it is registered and listed in the leading Web search engines and directories. Tools such as [Submit It!](#) or any of the many others now available (see [Yahoo's listing in this area](#), can make online submission to search engines a quick and easy task. All of the leading search engines and Internet portals must be targeted, although the issue of context is again very important.



Your gateway needs to be included in search engines like [Alta Vista](#) and [Yahoo](#), as many people use these as their starting points when searching the Web. However, subject-specific, geographically limited and specialist search engines should also be considered. Is there a local search engine that your users may frequent? If so, then registering your gateway with the site could pay off. If you can get listed on the most popular site (in terms of your target audience), then the relevance of the materials will be high and so the chances of people following links to your site are much greater.

Getting the most from search engines requires the use of metadata in your information gateway Web pages. This will not be a problem for a metadata expert!

### CROSS REFERENCE

[Metadata formats](#)

### Mailing lists and newsgroups

Many people are now familiar with the benefits of newsgroups and mailing lists and their power to contact large numbers of people with a specific interest. These can be excellent tools via which larger numbers of target users can be contacted. All it takes is an email or a news posting and your gateway's latest features can be publicised to hundreds or thousands of people. It also only takes one inappropriate message to alienate lots of potential users. Be careful of sending too many or inappropriate messages to newsgroups or mailing lists, as promotion can easily turn to spam.

### Face-to-face activities

The final area that should be considered in terms of promotion and publicity is that of face-to-face contact with potential users. Clearly, the effective way to do this is at large gatherings of potential users such as conferences and workshops. A presentation, paper or demonstration at a leading conference which will be well attended by potential users can communicate directly with a large group of users who may be influential. Running workshops for sections of your user community, especially for those who are themselves involved in training, can have similar results and is covered in more detail in the training and skills development chapter.

### EXAMPLE

#### Examples: Gateway presentations

- **Biz/ed:** *EBEA (Economics and Business Educators Association) Annual Conference*  
One of the key groups of users targeted by Biz/ed comprises UK economics and business school teachers. Over the years Biz/ed has given a number of presentations to the EBEA annual conference, as well as running an information stand about the gateway; presentation topics have included 'An introduction to Biz/ed' and 'Using the Internet in GNVQ Business'. All the presentations have served to highlight the Biz/ed information gateway directly to key users.
- **SOSIG:** *IRISS 98: Internet Research and Information for Social Scientists 1998*  
The IRISS conference was a leading conference for social scientists interested in using the Internet in their teaching and research. Debra Hiom from SOSIG did a presentation 'The Social Science Information Gateway: Putting Theory into Practice' which detailed many of the uses and strengths of the SOSIG information gateway.

### Are all types of publicity worth while?

---

The old saying that all publicity is good publicity probably has some truth, even when talking about information gateways. Any promotion and publicity that raises the profile of your gateway in your target community should be considered a good thing. Of course being voted the worst Web site by your user community should probably be avoided, but it may bring you a few curious visitors!

### How can I best target a limited budget?

---

The issue of how best to target a limited budget really depends on the makeup of your user community. If you have a wide user community, then you will have problems in targeting users. A well defined user community can often be more easily targeted as its members appear in concentrated groups or areas (within certain University Departments or organisations). A good example of this is the SOSIG user community, which can be relatively easily targeted via UK higher education social science departments.

### Are there any failsafe methods for successful publicity and promotion?

---

Unfortunately the answer to this is no. However, some of the existing gateways have demonstrated that certain techniques can be very cost effective; training trainers within your user community can produce very good results (e.g. Biz/ed) and well-placed publicity leaflets and posters in HEI libraries and departments can also communicate with large numbers of target users (as has happened in the cases of SOSIG and NMM Port).

Your user community should be carefully characterised before any expensive promotional activities are embarked upon. Identify your users carefully and your promotional activities will be much more likely to succeed.

### How can you retain the interest of your users?

---

Once you have persuaded potential users to look at your gateway, you would like them to come back to it. A well-designed gateway which fulfils the expectations of its users will encourage them to return, but publicity can also help them to keep the gateway in mind.

An email list can be a useful way of conveying information about developments in your gateway to interested users. Such a list has been run successfully for the SOSIG information gateway.

### References

---

Alta Vista, <http://www.altavista.com/>

Biz/ed, <http://www.bized.ac.uk/>

NMM Port, <http://www.port.nmm.ac.uk/>

OMNI, <http://www.omni.ac.uk/>

SOSIG, <http://www.sosig.ac.uk>

Submit It!, <http://www.submitit.com/>

D. Hiom, 'Around the table: Social scientists have their own favourite places on the Web', *Ariadne 9* (May 1997).

D. Hiom, 'SOSIG: Providing access to internet information', *Laser Link* (Autumn 1998).

J. Kirriemuir, 'A report on the third annual OMNI seminar: A cure for information overload', *CTICM Update 8:2* (December 1997).

C. Sladen, 'Ethical Business', *Business Review* (April 1998).

C. Sladen, 'Mergers and Take-overs', *Business Education Today* (May/June 1998).

### Credits

---

Chapter author: [Martin Belcher](#), [Lesly Huxley](#)

With contributions from: Sarah Ashton (NMM Port), Kate Sharp (Biz/ed), Debra Hiom and Emma Place (SOSIG).

## 2.9. User interface design

### In this chapter...

---

- identifying your target users - who are the potential users of your gateway?
- user consultation - asking your users about their wants, needs, likes and dislikes
- task analysis - what kind of tasks are they going to carry out using your gateway?
- usability and accessibility - what do these often-used terms really mean?
- Web design issues - Web design = information gateway design?
- developing a user interface design specification

### Introduction

---

This chapter looks at the general user interface issues which should be considered when planning the development of an information gateway or when looking at the modification of an existing gateway. Many of the issues discussed apply to all online services and Web sites, so they can be re-used outside the information gateway arena.

The importance of good user interface design:

- information gateways need to be usable
- the user interface to an information gateway is the design employed in the Web pages of the gateway
- good Web page design can significantly increase the ease with which users can complete tasks, i.e. it increases usability
- users who can't complete tasks are frustrated users; frustrated users don't come back
- users who complete tasks are happy users; happy users come back to a Web site and often tell their friends and colleagues about a great site/information gateway

### Gateways in context

Information gateways are really just value-added Web sites. This statement is not meant to belittle the importance of information gateways (far from it!); rather it is meant to highlight the fact that they have many similarities with Web sites in general. For all that is said about the Web being an interactive medium and an empowering tool from the user's perspective, there is one small point often overlooked. This is that the only way a user can interact with even the most advanced Web site is via the user interface. The user interface is simply what the user sees on the screen through their browser. If what they see is hard to understand or difficult to use, then the vast majority of users will never make it to the real content or value-added features of the Web site. It doesn't matter how good the information on your Web site is - if the user can't access the information, they will go elsewhere.

### Frustrated users

How many times have you visited 'great looking' Web sites and found them difficult to use, often so difficult that you have given up and gone elsewhere?

Poor user interface design can hide even the most powerful and useful Web sites from all but the most advanced and patient users. Web site developers (including information gateway developers) have to consider seriously the issues of user interface implementation. A poor user interface will mean low usage of the site and its ultimate failure. The failure of Web sites is often due to their designers' not considering their users and designing with the assumption of too much technical knowledge.

It should always be remembered that, by being in the position of developing or even just considering the development of an information gateway, you are probably in the category of an advanced user. You may not be as advanced as the system administrator or 'techie' in your organisation, but compared to the average man in the street you are an expert! Never overestimate the skills of your users, unless you have direct evidence on which to base your judgements.

## Background

---

### Definitions:

- **user interface:** the means of communication between a human user and a computer system (in this case a Web site). A wider definition could be the means of interaction between a human being and any object
- **usability:** the degree of ease with which human beings can interact with an object, in particular a computer system
- **accessibility:** the characteristics of Web content and whether or not it is accessible to people with disabilities

The science of user interface design, usability and accessibility has its origins in software development and general engineering. Many of the things we take for granted have been through a lengthy process of user interface design and development. Generally we don't notice interface design unless there is a problem, resulting either from poor design or from our attempting to use an object for something other than the purpose for which it was designed.

### EXAMPLE

#### Example of user interface design

Have you ever thought about the user interface design of a pair of scissors?

Scissors have actually been carefully designed for a specific range of tasks. Their design isn't really an issue unless you stress test a pair: if you are left-handed and try to use a pair of right-handed scissors, you immediately see the user interface design limitations. Ask a left-handed colleague to explain or try using a left-handed pair of scissors - you will wonder what the designers are playing at!

As mentioned above, most manufactured objects have some degree of user interface evolution and redesign involved in their development. Many household objects have been around for many years and so have the benefit of gradual development (scissors have been with us for hundreds of years). Unfortunately software design and development has been around for a much shorter period of time, and Web site design even less. The end result is that the usability of computer systems and Web sites is not completely understood or, in some cases, even recognised.

However, in order to develop successful information gateways you must consider the user interface design carefully and thoroughly. Without sufficient effort being put into this area you may be set for failure from the outset.



#### Useful resources

The following resources are extremely good and highly recommended as excellent introductions and background information on usability and user interface design (even if they do come from a single source):

- [Differences between print design and Web design](#)
- [How users read on the Web](#)
- [Be succinct! \(writing for the Web\)](#)
- [The top ten new mistakes of Web site design](#)

So what issues do I need to consider in order to develop a successful user interface?

## Identify your target users

---

It may sound obvious, but you can't really start thinking about the design of a user interface until your users have been identified and characterised. User identification is important in other aspects of the development of an information gateway (scope policy, gateway aims and objectives, planning an information gateway project), so that the question of who the target users are should have already been considered.

Different groups of users will vary in their characteristics. Wherever possible, you should try and include as large a range of users as you can, but think carefully about designing for everyone. If your target users have slightly different characteristics from the general public, then you have to prioritise which characteristics you wish to address.

When you are identifying your users, a minimum set of characteristics to consider might be:

- location of users (organisational and geographical)
- subject knowledge (educational level)
- IT literacy/technological experience (do not overestimate)
- access to technology and network connectivity
- physical attributes (colour blindness, age, disabilities)

Some of these characteristics can be obtained from correlation with general population characteristics, while others must be uniquely researched.

## User consultation

---

Once you have identified who your target users are, you may wish to consider having some degree of user consultation. Ideally, this would have been a part of the general development of the information gateway project/idea. The value of user consultation should not be underestimated. A few relatively simple techniques of user consultation can produce extremely powerful data which can influence the development of a user interface.

In the past, user consultation was often not considered, as it was thought to be time-consuming, difficult and contrary to the prevailing culture of 'we know best'. All these issues can be addressed by adopting a number of techniques that are simple to implement, low cost and able to provide convincing evidence of the power of user consultation.

### Questionnaires and surveys

The development and implementation of a simple questionnaire and survey of potential users can also produce important information. Selecting the people to be surveyed is important (so as not to build any bias into data collected), as is the careful wording and development of the questions that are being asked. Again, you would be well advised to consult some of the leading literature or any in-house experts.



### TIPS

#### Useful resources

The following resource is an excellent starting place for further information on conducting questionnaires and surveys:

- Questionnaire Design, Interviewing and Attitude Measurement. A.N. Oppenheim. 1992

A questionnaire is a good method of sorting and selecting the attendees for the next area of user consultation, focus groups.

### Focus groups

The focus group is a simple concept, although easy to implement wrongly. The basic idea is to get some target users in a room, ask them questions about the proposed information gateway and collect their feedback on your questions and ideas. Suggestions and problems can often come to light from a simple focus group discussion. Participants can highlight areas that have never been

considered by people too closely involved in the project.

Focus groups do need to be run with care, as they can often produce misleading information and are easy to run badly (for example, it is very easy for the person running the focus group to lead the answers as well as the questions!). The science of focus groups has its own extensive literature and it would be worth consulting one or two of the leading publications in this area.



### Useful resources

The following resources are excellent starting places for further information on running focus groups:

- The Focus Group: A Strategic Guide to Organising, Conducting and Analysing the Focus Group Interview. Jane Farley Templeton. 1994
- Focus Groups : A Practical Guide for Applied Research. Richard A. Krueger. 1994
- Focus Groups: A Step-By-Step Guide. Gloria E Bader, Catherine A. Rossi. 1998

### User consultation warning

Although user consultation is an essential part of any detailed user interface implementation project, it must be treated with some caution; there can sometimes be a marked difference between what users say they want and what users actually use. This is particularly true when complex features have been developed and implemented; user tracking and logging may show that very few people use the features. Some of the lack of use may be due to usability problems and some may be because users just do not want to use complex features.

User consultation should ideally go hand in hand with user tracking and logging of behaviour. Much user behaviour tends to be common across the board and it would be extremely useful if the information gateway community actively shared such information.

## EXAMPLE

### Guerrilla HCI

The term 'Guerrilla HCI' was coined by Jacob Nielsen in the field of software design and development. His basic premise is that software projects often fail to achieve their full potential because of the lack of user consultation, which is not considered because of the perceived high costs. Nielsen developed the idea of relatively low-cost user consultation and, although not directly related to Web site development, there are many useful issues raised in his publications on these issues.

The following document contains many insights and suggestions which may be directly applicable to gateway projects that are interested in a degree of user consultation and usability testing, but are not operating on a large budget:

- [Jacob Nielsen: Guerrilla HCI: Using Discount Usability Engineering to Penetrate the Intimidation Barrier](#)

### Task analysis

---

The outcome of any user consultation and/or user identification should be an understanding of the needs and requirements of the user community and an idea of what kind of tasks the average user is going to want to be able to perform. The ultimate aim of any user consultation should be to inform the gateway developers about the users' needs. Do the characteristics of the user community mean that they have any unique needs? For example, are they all on very slow network connections and only using text browsers, or are they all based in Higher Education Institutions (HEIs) and therefore have access to fast network connections?

The development of a description of and set of characteristics for a typical user will help to determine a set of user needs. This in turn will provide evidence to feed into a user interface requirements specification.

Information on task analysis can also be obtained from user consultation; getting participants in a focus group to discuss the kinds of tasks they might like to perform while using a gateway may help to decide the level of priority tasks should be given within the overall user interface design. Are the users' requirements, as described by the users, the same as those determined by the gateway developers? They should be similar but it is unlikely that they are the same.



## REMEMBER

### Existing gateways: user consultation?

If your gateway is already up and running, then user needs analysis and task analysis can significantly help you to improve the user interface design. User consultation and usage log analysis can help to refine an existing gateway, the better to meet user needs and expectations. Are users still using the gateway in the way originally envisaged? Asking them may reveal this, and looking at logs of how they use the current site can provide even more information. If your gateway offers browsing and searching, which one is being most heavily used? If there are significant patterns emerging from any data that you analyse, is a revised user interface called for?

## Usability and accessibility

---

Usability and accessibility often go hand in hand; if a Web site is difficult to use then it may become inaccessible, as users cannot get to the information that they want. Making something more accessible often makes it more usable for all users. Designing for maximum accessibility helps designers to focus on users and content rather than on 'flashy' design issues.

But accessibility also needs to be considered with regard to people with disabilities and giving equality of access to a Web site or information gateway. By making sure that a Web site is accessible to as wide an audience as possible you also necessarily increase the usability of the site. Catering for disabled accessibility may be something that a gateway would like to do or something that it is legally required to do (Hotwired '[Sites Must Retool for Disabled](#)'). In either case the issues need to be looked at and carefully considered. More detailed information on accessibility is contained in the Usability and Accessibility chapter.



## CROSS REFERENCE

[Accessibility and usability](#)

## General Web design issues

---

Web design is a science in itself and there are countless books and online resources that offer extensive advice in this area. A few key issues should be considered when designing:

- always design for your users and not the person running/funding the project
- be aware of and implement some degree of usability and accessibility standards
- avoid proprietary technologies, unless a significant proportion of your user community demand them
- try to use innovative and exciting Web design but don't overdo things

## Developing a user interface requirements specification

---

Before any implementation of a user interface begins, a detailed user interface requirements specification should be developed. The document should state the characteristics of the target users and for which tasks they are going to use the information gateway. There should also be a list of user interface priorities, with clear indications as to what is an essential requirement and what is desirable. Without such a prioritised list, it is difficult to decide where staff effort should be spent in user interface development. Unless there is an order of priority, if only some things are implemented, there will be no guarantee that they will be important in terms of usability and accessibility.

A good example of a well structured and well planned requirements specification is the [W3C Web Accessibility Initiative Standard \(WAI\)](#) and in particular the [List of Checkpoints for Web Content Accessibility Guidelines 1.0](#).

The document is useful in that it provides an excellent example of how to present a requirements specification document in an easy to understand and usable format. Additionally, it presents the definitive set of guidelines on how to implement a Web site of any description which has accessibility at its core. The document should be consulted by developers of all information gateways, current and planned.

## Case Studies

---

### EXAMPLE

#### Biz/ed usability audit

The Biz/ed information gateway was one of the early information gateways, originally launched in January 1996. The user interface that was developed then reflected the style and knowledge of the general Web technologies available at the time. Several minor redesigns were implemented as a result of internal changes to the site and general Web developments. The end result was something that looked acceptable and seemed to work.



Screen shot of Biz/ed homepage 08/07/96

In late 1998 it was decided that there would be some formal user consultation to see how users were using the site and to see whether there had been any changes over time. Analysis of the Web site user and search term logs indicated that people were using the site differently from the way in which they had used it earlier.

A series of focus groups and usability testing sessions were conducted over several months, to ascertain what it was that users liked about the information gateway as it then stood. Biz/ed also wanted to see if some proposed changes to the site would be popular. The outcome of the user consultation was that some changes to the site were implemented as planned, some were modified and some left out altogether. The participants in the focus groups and usability testing sessions also contributed significantly to the new user interface design. Simple techniques such as naming and grouping, user tracking, focus group issue investigation and task completion analysis were all employed to provide data for the gateway redesign.







Screen shot of Biz/ed homepage 11/07/99



### Cost of user testing a Web site

It takes 39 hours to test a Web site for usability the first time you try. This time estimate includes planning the test, defining test tasks, recruiting test users, conducting a test with five users, analysing the results, and writing the report. With experience, Web user tests can be completed in two working days.

- Jacob Nielsen: [Cost of user testing a Website](#)

## Glossary

---

**Accessibility** - the characteristics of Web content and whether or not it is accessible to people with disabilities

**Guerrilla HCI** - Term coined by Jacob Nielsen to describe the rationale behind discount usability engineering and how to put it into practice. Further information can be found at [http://www.useit.com/papers/guerrilla\\_hci.html](http://www.useit.com/papers/guerrilla_hci.html)

**HCI** - Human Computer Interaction

**HEI** - Higher Education Institution

**Heuristic evaluation** - Heuristic evaluation is a discount usability engineering method for quick, cheap and easy evaluation of a user interface design. Further information is available at <http://www.useit.com/papers/heuristic/>

**Usability** - the degree of ease with which human beings can interact with an object, in particular a computer system

**WAI** - Web Accessibility Initiative Standard

## References

---

Biz/ed,  
<http://www.bized.ac.uk/>

G. E. Bader & C. A. Rossi, *Focus Groups: A Step-By-Step Guide* (1998).

R. A. Krueger, *Focus Groups : A Practical Guide for Applied Research* (1994).

J. Nielsen, *Cost of user testing a Website*  
<http://www.useit.com/alertbox/980503.html>

J. Nielsen, *Guerrilla HCI*  
[http://www.useit.com/papers/guerrilla\\_hci.html](http://www.useit.com/papers/guerrilla_hci.html)

J. Nielsen, *Differences between print design and Web design*  
<http://www.useit.com/alertbox/990124.html>

J. Nielsen, *How users read on the Web*  
<http://www.useit.com/alertbox/9710a.html>

J. Nielsen, *Be succinct! (writing for the Web)*  
<http://www.useit.com/alertbox/9703b.html>

J. Nielsen, *The top ten new mistakes of Web site design*

<http://www.useit.com/alertbox/990530.html>

A.N. Oppenheim, *Questionnaire Design, Interviewing and Attitude Measurement* (1992).

J. F. Templeton, *The Focus Group: A Strategic Guide to Organising, Conducting and Analysing the Focus Group Interview* (1994).

W3C, *List of Checkpoints for Web Content Accessibility Guidelines 1.0*

<http://www.w3.org/TR/WAI-WEBCONTENT/checkpoint-list.html>

W3C, *Web Accessibility Initiative Standard (WAI)*

<http://www.w3.org/TR/WAI-WEBCONTENT/>

## Credits

---

Chapter author: [Martin Belcher](#), [Phil Cross](#)

With contributions from: Jan Chipchase

## 2.10. Integration of robot and manual indexes

### In this chapter...

---

- This chapter will be available spring 2000, when the handbook will be revised and updated.

## 2.11. Distributed cataloguing

### In this chapter...

---

- advantages of distributed cataloguing
- distributed cataloguing models
- management issues
- a case study: SOSIG
- examples of distributed cataloguing

## Introduction

---

This chapter introduces the concept of distributed cataloguing and the potential for working collaboratively across the Internet. It looks at some of the human issues involved in distributing cataloguing effort, presents some models currently in use within information gateways and in particular looks at the experiences of SOSIG in employing a distributed model. Some further examples of distributed cataloguing models are also presented.

Because of the open nature of the Web there is considerable potential for distributed collaborative cataloguing of networked resources. Information gateways can be built by teams of staff who are geographically dispersed but who can add resources to a database from their desktops via the WWW. This chapter concentrates mainly on issues surrounding distributed cataloguing into a central database. However, an additional or even complementary model is that of collaborative work with other gateways (see the chapter on co-operation for more details).

### CROSS REFERENCE

[Co-operation between gateways](#)

### Why would an information gateway want to consider distributed cataloguing?

Distributing the cataloguing effort allows you potentially to share the responsibility with a number of organisations or participants and to maximise the coverage of the collection. In particular it allows gateways to:

- locate cataloguing effort within centres of subject expertise
- locate cataloguing effort within centres of geographical knowledge
- provide access to staff with a variety of language capabilities, enabling the development of multilingual gateways
- aid economies of scale

### Models for distributed cataloguing

---

There are numerous cataloguing models currently being employed by information gateways. The main contrast is that of the use of paid versus voluntary effort. However, even within this broad division there are several approaches, e.g.:

- networks of volunteers
- institutional commitment to provide staff effort as part of their main duties
- paid staff
- a mixture of paid staff and volunteers

And within these organisational setups there are various ways of assigning roles and responsibilities. These range from allowing members of the team to have full responsibilities and access to the database to a very defined division of labour between selecting, evaluating and cataloguing resources.

DESIRE 1 held a training workshop on the Distributed Cataloguing Model in 1997, which brought together staff from a number of European information gateways to share experiences of their models and the tools, training materials and methods of delivery to support them. A report summarising the outcome of the workshop can be found at:

<http://www.desire.org/results/training/D8-2af.html>

### Management issues

---

There are a number of issues to consider when setting up a distributed cataloguing system.

#### Recruitment

One of the most crucial issues for gateways is recruiting the right staff to work on the catalogue. The core skills of resource selection and cataloguing make librarians ideally placed to assume the role, as they have the training and the expertise required. However, academic subject experts or others with the appropriate subject knowledge may also be valuable. It is also important to bear in mind that as well as subject knowledge a fair degree of expertise in use of the Internet is also necessary and that these two skills are not always found together.

As well as deciding on the type of person required, gateways will also need to consider the best approach to finding and recruiting these people. Putting out a general call for staff will usually result in receiving replies from enthusiastic individuals who are keen to do this sort of work. However, they may have difficulties in getting the support they need to do this from their institution or place of work. Conversely, going through the institution will ensure commitment from the top down but may not result in the ideal candidates being selected from within the institution.

A key decision is whether the staff will be volunteers, will include the work as part of their jobs or be paid for their contributions. Paid staff will enable gateways to set and work to targets allowing for the development of the gateway to planned and monitored. With voluntary effort gateways are relying on the goodwill of the people concerned and the ability to fit these duties around their main jobs and activities. It is quite possible that there will be very little return for the considerable investment made in training and development. Perhaps the ideal situation is to have staff who are supported by their institutions to incorporate the role into their day-to-day work. Ensuring that paid staff have protected time to carry out their gateway duties may also be an issue; it is possible that external staff have been given this additional role on top of their existing work and will find it difficult to cope with both. Good communication between the central and distributed staff can help to prevent these problems arising.



[Subject indexing and classification](#)

## Support tools and mechanisms

Gateways need to develop a system for staff to be able to remotely recommend or catalogue into the system. Again, various methods are used by gateways; these range from emailing details of resources to central staff to Web based cataloguing systems such as ROADS.

## Training

Training staff to contribute to the gateway is essential. They will require training in:

- selection of resources
- cataloguing and classification

Ideally this training would take place as a face-to-face workshop, although, given the possibility of contributors being located around the world, training could also take place through distance learning via email and the Web.

## Documentation

Whether training is conducted remotely or face-to-face, extensive documentation is required to support the work of the staff. Various approaches are being used by existing gateways. Some have printed handbooks with all the information required; others have set up administration centres on the Web with online documentation and support.

## Monitoring and support

Perhaps one of the greatest drawbacks of running a distributed team is dealing with the problems of working remotely. The job requires that staff should be self-motivated, yet it is very easy for staff to feel isolated without the advice and support of colleagues around them. A geographically dispersed team will rely heavily on remote communication through one-to-one email contact, use of mailing lists and Web conferencing systems for 'virtual meetings'.

## A case study: SOSIG

---

SOSIG has successfully employed a distributed team of subject experts (known as Section Editors) for the past two years. Subject librarians from ten UK universities were appointed to select, evaluate and catalogue resources for the SOSIG catalogue. Each Section Editor is given responsibility for developing a subject area on the gateway. In some cases the Section Editors' roles are shared between two or more people at an institution, but total effort does not exceed more than one day per week.

A one-day workshop was held at the start of the project to train the staff on all aspects of working on an information gateway. This included:

- introduction to the Scope Policy of SOSIG (this stipulates the audience and type of information to be included in the gateway)
- finding resources on the Web
- selecting and evaluating resources
- cataloguing resources via the Web (including cataloguing rules)
- introduction to SOSIG's Collection Management Policy (including guidelines on deselecting resources)

### CROSS REFERENCE

[Quality selection](#), [Collection management](#)

Prior to the workshop an online administration centre was set up, which included all the tools and guidelines required to catalogue resources for the gateway. After the workshop, additional support was offered through email contact with the core staff. This one-to-one contact was initially very important as the Section Editors had a very steep learning curve to ascend. The geographical distances between the staff meant that they were very reliant on email as a means of virtual support and assistance. As the Section Editors have direct access to the live database to begin with, all of the work submitted had to be checked centrally and any errors corrected and/or reported back to the appropriate Editor. This put a very high overhead on central effort for the first few

months of the scheme; however, this requirement diminished gradually and now only random checks are made on the records.

In addition to the Section Editors, SOSIG also has a number of European Correspondents. Correspondents are academics or librarians who have volunteered to submit new resources on an informal but regular basis. Correspondents have access to online training and support materials but they do not catalogue directly into the database; rather they are responsible for selecting resources and submitting the suggestions to the central team through an online form.

The responsibilities and duties for the gateway can be represented visually in two ways:

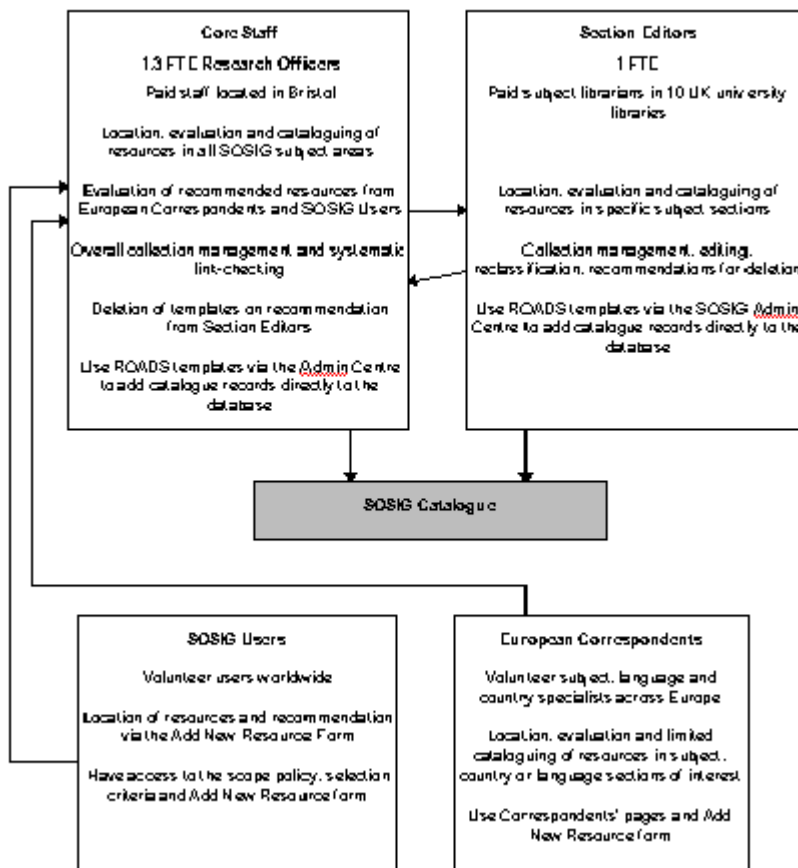


Figure 1: Workflow

	CORE STAFF	SECTION EDITORS	CORRESPONDENTS	SOSIG USERS
PAID				
LOCATE RESOURCES				
LOCATE NON-ENGLISH RESOURCES				
GENERATE RESOURCE DESCRIPTION & KEYWORDS				
GENERATE FULL CATALOGUE RECORD				
CHECK AND EDIT RECORDS				
LINK CHECKING				
DELETE RECORDS				
COLLECTION MANAGEMENT				
HANDLE USER RECOMMENDATIONS				

Figure 2: Tasks and responsibilities

There have been various general lessons learnt in the process of establishing this distributed approach as a result of other attempts by SOSIG to encourage distributed input, which may be relevant to other gateways. These are:

- an institutional commitment, backed by a financial arrangement, is a far more reliable way of establishing a broad range of participation than using volunteers or making financial arrangements with individuals
- such collaborations require a great deal of co-ordinating and supporting effort from the service centre, including training, responding to queries and general reassurance as well as monitoring and encouraging effort
- an essential ingredient has been the personal contact between the Section Editors; bringing them all together regularly for information-sharing and morale-boosting sessions has noticeably improved quantity and quality of results. Even though these face-to-face sessions are relatively expensive exercises, they have been well worth while.

#### E X A M P L E

#### Other examples of distributed cataloguing models

##### DutchESS

DutchESS (The Dutch Electronic Subject Service in the Netherlands) has a number of volunteer subject specialists from university libraries around the Netherlands. The subject specialists select resources and submit them to a local editor who checks the resources and edits the catalogue descriptions as appropriate. The local editors feed resources to DutchESS, where they are entered into a database. Face-to-face training for the subject librarians has been conducted. Interestingly, in this model the subject librarians involved work on this gateway as part of their day-to-day library work.

For more information see: <http://www.konbib.nl/dutchess/docs/info.html#8>

##### EELS

EELS (Engineering Electronic Library in Sweden) is an engineering subject gateway. EELS has ten Section Editors from university libraries around Sweden who volunteer to submit resources to the database. They catalogue resources directly into the database and are also able to delete records. The Section Editors receive face-to-face training in the key skills.

For more information see: <http://www.ub2.lu.se/eel/about.html>

##### EEVL

EEVL (Edinburgh Engineering Virtual Library in the UK) is a subject gateway to engineering information on the Internet. They have had a distributed team of academic librarians attached to the service from its inception. These librarians work voluntarily, but are part of the EEVL project consortium and so have a stake in the project. As such, they have been heavily involved in creating the selection and cataloguing process, so they have not required formal training, but they do have a printed procedures manual and regular meetings to discuss policy. The librarians can add records directly to the database, but these are checked by central staff before they are made publicly available on the gateway.

For more information see <http://www.eevl.ac.uk/volunt.html>

##### Friends of ADAM

ADAM (Art, Design, Architecture and Media Information Gateway in the UK) has created the 'Friends of ADAM' system. The Friends are a volunteer and support network from the arts and media community recruited through email, web publicity, and conferences and events. The system involves accredited online training in three areas:

- evaluation
- nomination
- cataloguing

At the end of the training period volunteers are issued with a certificate of competence and can

then assume different levels of responsibility in the service. Those gaining a certificate in evaluation and nomination feed their suggestions to the central team at ADAM who catalogue the resources into the database. They also assist in evaluating suggestions sent to ADAM by members of the public. Those gaining a certificate in cataloguing can create new resource descriptions which are checked by team members before being added to the central database.

For more information see: <http://adam.ac.uk/friends/>

### **[Länkskafferiet \(Link Larder\)](#)**

The Link Larder is a database for educational use and is intended as a pedagogical aid for Swedish pupils, especially those between 10 and 15 years of age, in their search for useful information on the Internet. All the web sites are selected, quality assessed and described by eight subject editors.

For more information see: [http://länkskafferiet.skolverket.se/information/brief\\_presentation.html](http://länkskafferiet.skolverket.se/information/brief_presentation.html)

## **Recommendations**

---

There is great potential for distributed cataloguing systems, as they open up the possibility of national or international strategies. They also provide a successful model for involving the library community in Internet resource discovery. Existing gateways have invested effort in developing systems that support the work of distributed teams, so that a librarian can work on a gateway from anywhere in the world as long as they have access to a networked PC and a Web browser. Distributed Internet cataloguing means that libraries can contribute to a shared service, rather than having each to build a local service. This is an efficient way of working, as it avoids duplicated effort and collaboration allows large-scale gateways with much better coverage to be developed.

Building and managing distributed teams is a challenge; there are a number of issues that need to be dealt with. In summary, some of these are:

- dealing with problems of distance and feelings of isolation - constant email contact and personal feedback on work is crucial to help alleviate these problems
- little control over individual work patterns - it is important to ensure that paid staff have protected time to carry out the work
- monitoring consistency between staff - this is much harder in a distributed environment, but providing clear and comprehensive documentation such as selection criteria and cataloguing rules can help

## **Glossary**

---

**ADAM** - Art, Design, Architecture and Media gateway  
**DutchESS** - Dutch Electronic Subject Service  
**EELS** - Engineering Electronic Library, Sweden  
**EEVL** - Edinburgh Engineering Virtual Library  
**SOSIG** - The Social Science Information Gateway

## **References**

---

*DutchESS Manual: handleiding voor vakspecialisten*, <http://www.konbib.nl/dutchess/manual/>

*EELS Project*, <http://www.ub2.lu.se/eel/about.html>

*EEVL*, <http://www.eevl.ac.uk/volunt.html>

*Friends of ADAM*, <http://www.adam.ac.uk/friends/>

*Länkskafferiet (Link Larder)*,  
[http://länkskafferiet.skolverket.se/information/brief\\_presentation.html](http://länkskafferiet.skolverket.se/information/brief_presentation.html)

SOSIG Correspondents Pages, <http://www.sosig.ac.uk/desire/ecorresp.html>

T. Hooper, L. Huxley & P. Hollands, *DESIRE: Subject-based training materials*  
<http://www.desire.org/results/training/D8-2af.html>

L. Huxley, '*DESIRE on Planet SOSIG: Training for the Distributed Internet Cataloguing Model*',  
*Ariadne 12* (1997).  
<http://www.ariadne.ac.uk/issue12/planet-sosig/>

E. Worsfold, '*Distributed and Part-Automated Cataloguing: A DESIRE Issues Paper*' (March 1998).  
<http://www.sosig.ac.uk/desire/cat/cataloguing.html>

## Credits

---

Chapter author: [Debra Hiom](#)

With contributions from: Rebecca Bradshaw, Roddy Macleod, Emma Place and Kate Sharp.

## 2.12. Multilingual issues

### In this chapter...

---

- providing a multilingual service
- technical issues
- interface issues
- metadata and cataloguing
- cross-language information retrieval

### Introduction

---

Gateways need to address the language needs of their audiences. Users may want to search a multilingual collection by using queries in one language or to retrieve documents in a number of specific languages, preferably also via an interface in the language of their choice. In some cases they may require some translation or summary in another language than that of the document. Ideally you should provide your audience with the language support it needs. In reality this will very likely be restricted, depending on the available technologies, the language skills of available staff involved in selection and cataloguing and cost considerations.

### Background

---

#### Multilinguality: praxis, trends and developments

There are two basic issues relating to multilingual access:

- the storing, processing and presentation of information in many languages (this is a question of enabling technology)
- multilingual search and retrieval

A lot of research has been going on in these areas for some time, especially in the retrieval of documents in languages other than that used for the query (cross-language information retrieval) (Oard, 1997). An overview of projects and demonstration systems can be viewed on the Web (compiled by Oard: <http://www.ee.umd.edu/medlab/mlir/systems.html>).

Nevertheless, existing gateways in general do not have much to offer yet in terms of multilingual support. Quite a few gateways - at least if they are not based in the UK or the US - do have a bilingual interface, usually the language of the country where the gateway is maintained and English, but more sophisticated facilities, such as multilingual search and/or browse support, are not often available. The main conclusion from a review conducted as part of the DESIRE I project in 1997 (Worsfold et al., 1997) was that there was considerable inconsistency in the way existing services deal with language issues. Not only did different gateways vary in their policies, there was also a lot of inconsistency within individual gateways. For example, titles are sometimes



displayed in the language of the resource, and sometimes only in English, and when resources are available in more than one language this is only sometimes mentioned. Some Internet search engines also offer a form of multilingual support, such as interfaces in various languages, localised search by country usually based on domain name, or automatic translation (such as Alta Vista's Babelfish, based on the Systran translation system). The services hardly ever describe the extent of their provisions in a detailed way, so it is difficult to assess what exactly they have to offer.

However, recent developments in the standardisation of metadata and resource description formats, electronic messaging and WWW technology can provide a solid basis for multilinguality in information gateways.

### **The European Multilingual Community**

The number of indigenous European languages, according to CEN TC 304, is 160. The Internet European multilingual community uses more than 30 languages, represented by many character sets with different repertoires and encodings. A property common to all of them is the use of the character-box (or glyph-box) representation or single-byte character sets (SBCS), i.e. each character uses one displayable position. In this they differ from other languages used outside Europe.

Most of the European languages use the Latin script, which consists of the 26 basic characters of the English alphabet (A through Z) in upper and lower case. Some languages, such as French, Spanish or Icelandic, need some additional characters, as well as a number of characters that are composed from the basic ones and the diacritical marks specified in a few basic ISO standards (such as ISO 6937). Fourteen diacritical marks, commonly called 'accent marks', which permit the support of nearly 200 diacritical combinations, complete the set for European Languages. [Demchenko]

The repertoires of the official European languages of the members of the European Union (EU) are specified in ISO 8859-1, while the repertoires of Central and Eastern European languages using the Latin alphabet are specified in ISO 8859-2. The Greek alphabet is specified in ISO 8859-7 and the Cyrillic alphabet used in Europe is specified in ISO 8859-5. The most widely used operating systems, such as UNIX and Microsoft Windows, use their own character set encoding (e.g. Windows Code Pages 1250-58 or ANS) for support of the European Languages including the Cyrillic languages (Russian, Ukrainian, Belorussian, Bulgarian, etc.) in CP1251 [Freed]. The de facto standards for mail and news exchange as well as for WWW information in Russian and Ukrainian speaking communities are KOI8-R (RFC 1489) and KOI8-U (RFC 2319). These different character set encodings implemented in different operating systems are the main source of problems in accessing Internet/WWW content with client software running on these systems.

## **Issues for Gateway Managers**

---

Gateway managers will be confronted with various choices relating to the language support of the service they want to provide. Those choices for monolingual or multilingual support present itself at many different levels:

1. Scope and selection policy.
2. Data presentation and resource description formats.
3. Metadata and cataloguing rules.
4. Searching and browsing.
5. The user interface.

### **1. Scope and selection policy**

---

Gateway managers will not be able to avoid language issues when trying to determine the scope and coverage of their service. They will need to decide whether to select all relevant documents, independently of their language, or to restrict the scope of the service to documents in one language or a number of specified languages. The following questions will have to be asked - and answered!

- will the service include resources written in more than one language, in any language or in a selection of languages?
- will the service include documents that require the use of Unicode or ISO 10646 character sets to support multiple languages and scripts in one single document, or it is possible to use single-byte character sets which normally contain characters from specific scripts together with the English alphabet/script (i.e. Latin 1, Latin 2, Cyrillic, Greek, Arabic, etc.)?

The choices made in this area directly determine the skills required of the staff responsible for

selecting and/or cataloguing the resources as well as the choice of the relevant authoring and access tools and software. For example, creating an information gateway that includes resources in all European languages would require input from a team who had mastered all those languages between them. If the cataloguing is done by a separate team, this team would also have to consist of people with various language skills. Not many gateways will be able to manage such broad coverage with an in-house team. A distributed model - as opposed to a centralised model - could offer a solution, by getting input from a multinational team, located in various countries, providing their input via the WWW. In this case a multilingual development framework needs to be implemented, based on standards in resource description formats (metadata) and information retrieval and exchange.

SOSIG provides an interesting case study of such a model. As the core team of SOSIG consisted of native speakers of English with no other language skills, SOSIG created a system whereby European correspondents suggest resources in a number of other languages to SOSIG staff. Problems with this approach are that the service is dependent on the goodwill of unpaid staff and that communication takes place (almost) exclusively in a virtual environment.

### CROSS REFERENCE

[Distributed cataloguing](#)



### REMEMBER

- the needs of your target audience
- technical features of the software underlying your service
- the skills of the staff responsible for selecting and/or cataloguing the resource
- the model for selection of resources (centralised or distributed), and (related to this) the available possibilities for ensuring the collaboration of staff or correspondents with the needed language skills
- the possibilities for the implementation of a multilingual development framework based on standards in resource description formats (metadata) and information retrieval and exchange as well as supporting development/authoring software.

## 2. Data presentation and resource description formats

---

A multilingual gateway would require the WWW software lying behind the gateway to cope with multilingual data handling, search, retrieval and display.

Existing standards and recommendations provide a framework for multilingual support in data communications and information resource description formats and metadata.

A model for multilingual support in Internet protocols and applications is defined in RFC 2130. It is implemented both in interactive applications, such as the WWW, and in non-interactive applications, such as electronic mail. Basic for interoperability in those applications is character set encoding (charset), which uses registered MIME (Multipurpose Internet Mail Extension) types, and language tagging, which uses registered language values or names according to RFC 1766 or ISO 639.

The HTTP protocol, on which the WWW is based, includes information about the type of the transferred information and the character encoding for text-based information, for example:

```
http-equiv="Content-Type" Content="text/html; charset=euc-jp"
```

The Content-Language entity header field describes the natural language(s) of the intended audience for the enclosed document:

```
http-equiv="Content-Type" Content-Language=se
```

If no Content-Language is specified, the default is that the content is intended for all language audiences.

It is also recommended to include information about the character encoding being used in the META information of the HTML document:

```
<META http-equiv="Content-Type" Content="text/html; charset=euc-jp">
```

Based on the exchange of information between client (browser) and server (HTTP Server) it is possible to provide character encoding and language negotiation between the information provider and the requester with regard to the accepted and preferred formats of the resources.

Recent developments in XML provide facilities for defining/labelling the language of the whole document, entity or item by including language attributes in the corresponding tag. For example:

```
<p xml:lang="en">The quick brown fox jumps over the lazy dog.</p>
<p xml:lang="en-GB">What colour is it?</p>
<p xml:lang="en-US">What color is it?</p>
<sp who="Faust" desc='leise' xml:lang="de">
</>Habe nun, ach! Philosophie,</>
</>Juristerei, und Medizin</>
</>und leider auch Theologie</>
</>durchaus studiert mit heißem Bemüh'n.</>
</sp>
```

Although the default XML Character Set Encodings are UTF-8 and UTF-16 (which are encodings for ISO 10646 or UNICODE), specific encodings for XML documents can be defined in the initial XML declaration for the whole document or entity (which can be regarded as a separately stored part of the whole document), for example:

```
<? xml encoding='UTF-8' ?>
<? xml encoding='ISO8859-1' ?>
```

Dublin Core, as a particular realisation of metadata resource description, provides possibilities for defining the language of the intellectual content of the resource, the record and the labelling language of particular fields by means of assigning language attributes to the relevant Dublin Core field.

### Examples

DC.Language Format	Field content language labeling/attribution.
<pre>&lt;meta name = "DC.Language" content = "en"&gt; &lt;meta name = "DC.Language" scheme = "rfc1766" content = "en"&gt; &lt;meta name = "DC.Language" scheme = "ISO639-2" content = "eng"&gt;  &lt;meta name = "DC.Language" scheme = "rfc1766" content = "en-US"&gt;  &lt;meta name = "DC.Language" content = "zh"&gt; &lt;meta name = "DC.Language" content = "ja"&gt; &lt;meta name = "DC.Language" content = "es"&gt; &lt;meta name = "DC.Language" content = "de"&gt;  &lt;meta name = "DC.Language" content = "german"&gt; &lt;meta name = "DC.Language" lang = "fr" content = "allemand"&gt;</pre>	<p>A work in Spanish may be assigned the following metadata:</p> <pre>&lt;meta name = "DC.Language" scheme = "rfc1766" content = "es"&gt; &lt;meta name = "DC.Title" lang = "es" content = "La Mesa Verde y la Silla Roja"&gt; &lt;meta name = "DC.Title" lang = "en" content = "The Green Table and the Red Chair"&gt;</pre>



- notwithstanding the future advent of total Unicode support in all system and application software, single-byte character sets will continue to be used as well for a long time. Your software should provide correct support and interoperability for both Unicode and single-byte encodings
- make sure that your workware (client and server software plus font supplement and cartridges) fully supports all working languages and used character sets at all layers of the multilinguality framework model
- configure your authoring tools (HTML and XML editors) in such a way that they insert metadata or attributes about language and character set encoding in the document. Don't forget to select proper parameters (character set encoding and Language) when you edit particular documents (in WYSIWYG HTML editors) or include this information when you use text editors for writing HTML documents
- when you provide multilingual information and use encodings other than US-ASCII or Latin 1 (ISO 8859-1) encoding, it is recommended to provide information as to where users can find or download the necessary fonts
- be sure that your HTTP server inserts correct information into the HTTP header. Note that different browsers may handle information about character set encoding in HTTP headers and metadata in the HTML document headers in different ways
- consider providing some basic training on multilingual issues for your core development staff

### 3. Metadata and cataloguing rules

---

If you enable the end-user to specify preferred languages, the search mechanism can return matches for resources that are in a language the user can read. Sometimes you also need to provide a selection of character set encodings to be correctly (i.e. in a readable way) displayed to the user. The latter is especially important for communities that use multiple character set encodings, i.e. charsets. Such selections can be provided as part of the client's browser and WWW server negotiation if they are defined by modern standards and supported by modern multilingual client/server software. For this to be possible the record must contain appropriate information. In other words, in order to be able to provide this option, some investment in multilingual development software/authoring tools and effort on the cataloguing side is necessary.

Traditional library practice is to create one record for one resource. On the Internet the question is what exactly constitutes a resource - the *granularity* issue. This is also relevant to language issues. Do you include only complete versions of the document, or do you also register parts of a site that are available in another language? If so, how substantial does the translated section have to be? A related issue is the problem of whether to create a separate record for each language version. For books this has been traditional practice; the translation of a book will get its own cataloguing record. For the Internet environment, it may be worth while to store information about different language versions in one record, as long as the fields relating to one version are linked in some way. It will be less labour-intensive to keep one record up to date, and there is no need to maintain a system of cross-references between language versions in order to keep track of different versions of one document.

Some services only mention the language of the resource in the free text description of the resource, not in a separate field, and often this is not very consistently done within one service. This means that the user may search on the word 'Swedish' in the description field and will thus find resources of which it is noted that they are 'Available in Swedish', but no separate formal support for searching on language will be possible, as the system has no properly encoded language information available on which to base such facilities.

To be properly handled by different software, language and character set encoding should be incorporated into metadata and resource description formats explicitly and in a correctly formalised way. The chosen metadata format will have to be able to accommodate this language information. For example both the Dublin Core element set and ROADS enable the storage of language information in a separate, repeatable element or field. ROADS allows the labelling of different variants of informative fields expressed in different languages. Dublin Core provides a mechanism to define the language of the content of a particular field as an attribute of this field. XML encoded DC (or RDF in general) can use an XML language attribute and character set encoding (\*\*on XML and DC, see above).

The metadata largely determine the search support that you will be able to provide. The more sophisticated your metadata set, and the more consistent the cataloguing practice, the more

advanced the information retrieval options you will be able to support. On the other hand, 'garbage in = garbage out'.

Two of the most widely used protocols for library and general network information retrieval, HTTP and Z39.50, allow language and character set encoding negotiation for each particular communication (HTTP-RFC2616, Z39.50-LANG). The general scheme for such negotiation is as follows:

- the requester or client (in the case of a WWW browser), sends a list of accepted character set encodings (charsets) and an accepted language priority list together with the URL/URI identifier
- the server/database returns the resource/document in the requested encoding and language, if it is explicitly labelled

Note that language and character set encoding negotiation that is provided on communication protocol level should normally coincide with correspondent information at document level (i.e. in the document itself). If this is not the case, the client can have problems in reading the requested information. It is the responsibility of the WWW server or database administrator to ensure that such a facility is implemented.

#### **multilingual issues in cataloguing:**

#### **CROSS REFERENCE**

#### [Cataloguing](#)

##### **1. Cataloguing of the title.**

Normally the title will be catalogued in the language of the resource. Titles for the same resource in other languages may be catalogued in an 'alternative title' field labelled with a language/variant label or attribute defining the language of the content. Some information gateways put alternative titles in the same field, separated by '=' or another symbol. It is recommended, however, to encode alternative titles in a separate field, with a language attribute or label, because this allows for more sophisticated handling of alternative titles in the search interface.



#### **REMEMBER**

- formalised encoding of alternative title information in the metadata format allows for more sophisticated handling of this information by the software
- defining a 'main' version and 'alternative' versions of a resource may cause problems, if it is not easy to determine what the main language of the resource is. For instance, what is the main title and what are the alternative titles for a Swiss resource, available in French, German and Italian?
- giving each language version its own record and cross-referencing the records means more maintenance
- when putting all the language information in one record, give all variants their own fields with attributes defining language
- that it is labour-intensive to have to check periodically whether other language versions of the same pages have been added
- what do you do with bits of a document that are in another language?
- do you want to translate the title of non-English resources into English?

##### **2. Language information in description/annotation.**

In the free-text description the language(s) in which the resource is available may be mentioned. This has some major disadvantages, because it is hard to guarantee consistency of practice and it does not offer a basis to specify language in the search process.

**REMEMBER**

- if you decide to adopt this approach, you could determine a default language to minimize effort. For instance, for resources available exclusively in English the language does not need to be mentioned, but an English page also available in French would get: 'Available in English and French.'
- when storing language information in the description field, structured search support for searching on the language of a resource cannot be provided
- it is almost impossible to check that the subject specialists/cataloguers consistently mention this information; the DESIRE review [Hiom et al.] indicated that this is not very consistently done

Another issue is the language of the descriptions themselves. There are several possibilities; the language of the description could be:

- the language of the resource it describes
- the language of the user interface and primary target audience of your service
- English as the Internet 'lingua franca'
- combinations of these, such as English and the language of your target audience

Descriptions in more than one language will of course multiply the necessary effort. A description in the language of the resource may be an option in a distributed model, with an international team of people without sufficient language skills in a common other language such as English, who select and catalogue resources in various languages. It may, however, be confusing to the user to be confronted with descriptions in various languages. Descriptions in a commonly used language such as English can give users information about documents in languages they can not read.

**3. A separate language field.**

The language of the resource may be in a separate field, preferably in a standardised format, e.g. ISO639 or RFC 1726. This facilitates search support for queries that specify the language of the resource. If different language versions are combined in one record, the alternative fields should be labelled so that they are linked to the title version that they belong to and the correct version of the title may be displayed to the user.

This practice is recommended instead of only mentioning the language(s) of the resource in a free text description.

**4. URIs.**

In the case where there is one record for different language versions, the URIs of all available language versions may be listed. In this case there should be some labelling of the URIs to link them to the title version to which they belong. Another option is to give just one URI, that of the home page, and let users choose their preferred language by using the language switch in the document. This will require less effort in creating the record and less maintenance; there can be only one possible 'dead link' instead of two or more. But, on the other hand, sometimes different language versions will be presented as equal, and it will be impossible to say which is the main version.

**REMEMBER**

- the language skills of the staff responsible for cataloguing the resources
- the way language is supported in the metadata format your are using (for instance Dublin Core, MARC, IAFA)
- the way language issues are handled in the cataloguing rules you use
- the search support you want to provide; these requirements must be met by the cataloguing format and rules

## 4. Searching and browsing

Cross-language information retrieval (CLIR) is the possibility of formulating queries in a natural language and retrieving documents in languages other than the language used for the query. The main approaches are defined (by Peters & Picchi, 1997) as:

1. Text translation via machine translation techniques.
2. Knowledge-based techniques - these involve the use of multilingual dictionaries, thesauri or general purpose ontologies.
3. Corpus-based techniques\*.

\*In this approach large collections of texts are analysed to extract the information needed to construct application-specific translation methods. This usually involves vector space and probabilistic techniques.

The first two approaches are the most relevant for Information Gateways:

### 1. Text translation via machine translation techniques

For cross-language information retrieval, machine translation of the documents does not seem to be the most realistic option, because of the costs (and the fact that some aspects of it, such as treatment of word order, are redundant for CLIR). More feasible is the translation of the query into the language(s) of the document. Retrieved documents may then be translated for the user, if required, a service that Alta Vista currently provides. It would be possible to add this service to an information gateway. Although results of machine translation are far from perfect, readers may prefer a flawed translation of a document they cannot read to none at all.

### 2. Knowledge-based techniques

First attempts involved matching the query to the document using machine-readable dictionaries, but the best results have been reached with thesaurus-based approaches. The drawback is that thesaurus construction and maintenance is expensive, and training is required for optimum usage. In the case of thesaurus-based controlled vocabulary indexing and searching, a set of monolingual thesauri is used which all map to a common system of concepts. Instead of the labour-intensive manual assignment of thesaurus terms by indexers, research is being carried out in the area of (semi-)automatic assignment of terms. Thesauri may also form the basis for more complex cross-language free text searching, where the query must be mapped to possible terms in the language (s) of the documents. ISO 5964 recognizes three approaches to the construction of multilingual thesauri:

1. Ab initio construction, i.e. the establishment of a new multilingual vocabulary without direct reference to the terms or structure of an existing thesaurus.
2. Translation of an existing monolingual thesaurus.
3. Reconciliation and merging of existing thesauri in two or more working languages.

#### CROSS REFERENCE

[Subject indexing and classification](#)



#### [EuroWordNet](#)

This project, which ran till June 1999, aimed to develop a general purpose multilingual ontology: a multilingual database, which represents basic semantic relations between words in various European languages, with Princeton WordNet1.5 as starting point. The basic principle is the construction of monolingual wordnets, which maintain language specific differences, which are mapped to a common top-ontology.

Although some gateways use thesauri for subject access (OMNI) or to provide the user with additional assistance in the choice of search terms (SOSIG), little or no use has been made by gateways of the potential of using a thesaurus for multilingual retrieval.

### 3. Classification schemes

If resources are classified using the numerical code from a classification scheme which is available

in more than one language, this enables language-independent searching as well as the possibility of offering a browsing structure in more than one language.

#### E X A M P L E

- [DutchESS](#) offers a browsing structure based on the Nederlandse Basisclassificatie which is available in Dutch and English. A (slightly different) German translation of the same scheme is also available, which would make it easy to add a German interface in the future
- [Jyväskylä Virtual Library](#) offers a browsing structure in Finnish and English (this does not apply to all sections of the distributed Finnish Virtual Library of which the Jyväskylä Virtual Library forms a part)

When choosing a classification scheme for your service, consider:

- in which languages the classification scheme is available
- whether it would be feasible to translate the scheme into another language in which it is not currently available but which you require for your service

#### 4. Keywords

Keywords may be added to the resource description in any language. In this case also a consistent policy may enhance retrieval possibilities. A number of options are possible:

- add keywords in the (primary) language of the service (user interface)
- add keywords in the language of the document
- add keywords in English as the Internet 'lingua franca'
- add keywords in a number of languages

Keywords may be chosen from an uncontrolled keyword list or from a controlled vocabulary; when available in more than one language this will provide opportunities for searching documents in various languages by means of a query in one language. The user should be made aware of the available options.

#### CROSS REFERENCE

[Subject indexing and classification](#)

### 5. The user interface

---

A monolingual user interface will probably be in the language of your primary audience or in a language familiar to a broad audience, such as English. The advantage of this is that it will require less effort to maintain, but you will exclude users who are not familiar with your chosen language. In the case of an academic audience, you may usually assume a certain proficiency in English, but a broader audience may not have those language skills. If the interface is in the national language only, this means that you narrow your target audience to one language community, dependent on the number of native speakers and others with a certain level of proficiency in that language.

Providing an interface in more than one language means that you will reach a broader audience, but you will have to put more effort in maintaining your service.

The target audience that you wish to serve will be of major importance when choosing the interface language(s). Another issue to consider is whether you are willing and able to match your multilingual interface with multilingual search support. For instance, if you provide a browsing structure based on a classification scheme which is available in one language only, do you want to put effort into translating the scheme into another language used in your interface?

In general users should be made aware of the consequences of the way they formulate their queries. This is easier said than done, if you want to avoid extensive help files or cluttered interfaces. For example: a simple query (all fields) in French may retrieve a document with the specified word in the title, but it will not result in any hits in the description field, if the language used for the description is English. As is well known, users are not very keen on reading help pages, so the search interface design should aim to present the language options in an clear and intuitive way.



 **CROSS REFERENCE**

[User interface design](#)

**REMEMBER**

- the expected language skills of your audience; do you aim to address a well defined language community or do you wish to provide for a broader audience?
- do you have staff with the necessary skills to translate the interface pages, or are you prepared to meet the extra cost of third party assistance (translation service)?
- are you willing and able to invest in extra creation and maintenance effort for your interface?
- are you willing and able to match your multilingual interface with multilingual browsing and/or search support?

**General conclusions**

multilinguality is a complex issue. Although a lot of technology has become available in recent years, many problems have yet to be solved. In most cases gateways will not be able to provide more than very basic facilities if they need to keep costs within acceptable limits. However, from the above it may be clear that putting some effort into making consistent choices - based on user needs - concerning such issues as scope and selection policy, metadata and cataloguing, classification and subject indexing, as well as regarding the use of the appropriate technologies, may enhance the language support you will be able to provide in your service; it will allow you to project a clearer picture to your users of what your gateway is about. Any extra facilities will have their costs, though, in terms of extra initial effort, maintenance, required skills of staff and so on, and it is up to you to decide whether user benefits outweigh necessary efforts to provide them.

**General recommendations**

- try to obtain knowledge about the language skills and needs of your audience
- aim at an integrated and consistent approach to language issues for your gateway.
  - Examples:
    - when your documents are in Danish only, it is probably not worth while to provide your users with a bilingual Danish/English interface
    - if you are not going to provide any multilingual search support, should you put effort into a bilingual or multilingual user interface?
    - if your cataloguing system can't handle Japanese, shouldn't you exclude documents in this language from the scope of your service?
    - consider the language skills of the staff responsible for selection and cataloguing when you develop the scope and selection policy of your service.
- try to balance requirements of effort against expected results and benefits of multilingual support for your users
- provide your users with information about your language policy, and integrate language related search options into your query interface design in a clear and unambiguous way

**Glossary**

**CEN** - European Committee for Standardisation  
**CLIR** - Cross Language Information Retrieval  
**CTE** - Content Transfer Encoding  
**DC** - Dublin Core  
**DutchESS** - Dutch Electronic Subject Service  
**IAB** - Internet Activities Board  
**IETF** - Internet Engineering Task Force  
**ISO** - International Standards Organization  
**MARC** - MACHine Readable Cataloguing. A family of formats based on ISO 2709 for the exchange of bibliographic and other related information in machine readable form.  
**MIME** - Multipurpose Internet Mail Extension  
**OMNI** - Organising Medical Networked Information (Medical gateway in the UK)  
**POSIX** - Portable Operating System Interface  
**SBCS** - single-byte character sets  
**SOSIG** - The Social Science Information Gateway  
**Unicode** - A universal 16-bit encoding for the scripts of the world's principal languages

**UCS** - Universal Character Set

**UTF** - UCS transformation formats - encodings for ISO 10646 or UNICODE

**XML** - Extensible Markup Language. A lightweight version of SGML designed for use on the Internet.

## References

---

DutchESS, <http://www.konbib.nl/dutchess/>

EuroWordNet, <http://www.hum.uva.nl/~ewn/>

Jyväskylä Virtual Library, <http://www.jyu.fi/library/virtuaalikirjasto/engroads.htm>

SOSIG, <http://www.sosig.ac.uk/>

Unicode Consortium, <http://www.unicode.org>

H. Alvestrand, *RFC 1766, Tags for the Identification of Languages (UNINETT, March 1995)*.  
<ftp://ftp.isi.edu/in-notes/rfc1766.txt>

G. Clavel et al., *CoBRA+ working group on multilingual subject access : Final report (Bern, 9th March 1999)*.  
<http://www.bl.uk/information/finrap3.html>

Y. Demchenko, *i18n and multilingual support in Internet mail Standards. Overview*.  
<http://www.terena.nl/multiling/>

Encoding Dublin Core Metadata in HTML (Internet Draft).  
<http://www.ietf.org/internet-drafts/draft-kunze-dchtml-01.txt>

Extensible Markup Language (XML) 1.0 (W3C Recommendation, 10 February 1998).  
<http://www.w3.org/TR/1998/REC-xml-19980210>

The ISO 8859 Character Sets  
<http://www.terena.nl/multiling/ml-docs/iso-8859.html>

ISO 639, 'Code for the representation of names of languages'.

ISO/IEC 10646-1:1993(E), 'Information technology - Universal Multiple-Octet Coded Character Set (UCS) -- Part 1: Architecture and Basic multilingual Plane' JTC1/SC2 (1993).

J. Knight, *Internationalization in the DESIRE project*  
<http://www.roads.lut.ac.uk/DESIRE/Desire18N.html>

D. W. Oard, 'Serving Users in Many Languages : Cross-Language Information Retrieval for Digital Libraries', *D-Lib Magazine* (December 1997).  
<http://www.dlib.org/dlib/december97/oard/12oard.html>

D. W. Oard, *Cross-Language Information Retrieval Resources (Overview)*.  
<http://www.ee.umd.edu/medlab/mlir/>

C. Peters, & E. Picchi, 'Across Languages, Across Cultures : Issues in multilinguality and Digital Libraries', *D-Lib Magazine* (May 1997).  
<http://www.dlib.org/dlib/may97/peters/05peters.html>

RFC 2413. *Dublin Core Metadata for Resource Discovery*  
<http://www.ietf.org/rfc/rfc2413.txt>

RFC 2616. *Hypertext Transfer Protocol -- HTTP/1.1*  
<http://www.ietf.org/rfc/rfc2616.txt>

The Unicode standard, version 2.0 (Unicode Consortium. Reading, Mass.: Addison-Wesley Developers Press, 1996).

C. Weider, C. Preston, K. Simonsen, H. Alvestrand, R. Atkinson, M. Crispin & P. Svanberg, *RFC 2130 - Report from the IAB Character Set Workshop (April 1997)*.  
<ftp://ftp.isi.edu/in-notes/rfc2130.txt>

E. Worsfold et al., *Developing multilingual subject gateways (An issues paper written as part of the DESIRE Cataloguing Project)*  
<http://www.sosig.ac.uk/desire/lang/language.html>

F. Yergeau, *RFC 2279 - UTF-8, a Transformation Format of Unicode and ISO 10646 (January 1998)*  
<ftp://ftp.isi.edu/in-notes/rfc2279.txt>

## Credits

---

Chapter author: [Yuri Demchenko](#), [Marianne Peereboom](#)

## 2.13. Co-operation between gateways

### In this chapter...

---

- strategic advantages of co-operation
- models for co-operation
- interoperability issues
- practical demonstrations of co-operative work
- key initiatives in gateway co-operation to date
- recommendations

### Introduction

---

The Internet offers great potential for co-operation between gateway services, since it allows geographically distributed databases and people to communicate with one other and to work together to build integrated services.

Co-operation between gateways is increasingly being seen as a strategy for:

- enhancing Internet resource discovery for end-users
- improving the efficiency and sustainability of gateway services

There are a number of different models for collaborative work, and, as gateways are still a relatively new type of information service, there is still much scope for exploring the potential of co-operation. Those running gateways should consider the benefits of, and opportunities for, co-operation with other gateways.

### Strategic advantages of co-operation

---

Why should a gateway consider co-operation with other gateways?

#### Enhancing Internet resource discovery for end-users

The development of a myriad of information gateways on the Web is, ironically, making it increasingly difficult for users to search the Internet effectively. Many gateways are claiming to offer a 'one-stop shop' for finding information and this may work for certain users; however, other users will benefit from searching more than one gateway. With lots of independent and uncoordinated gateways, this can involve making a series of searches in a number of services, all of which have different interfaces and ways of working. Not easy!

Collaboration can help gateways to offer integrated services for end-users. The advantages of this for users (depending on the co-operative model used) may include:

- access to far broader collections than any single gateway could offer, including high quality Internet resources on many subjects, from many countries, written in many languages
- access to a large number of metadata records via a single user-friendly interface

- the ability to locate new gateways that they may not have heard about
- the possibility of searching a selection of gateways simultaneously as opposed to one by one

### **Improving the efficiency and sustainability of gateway services**

As more organisations invest in building gateway services, more opportunity for collaborative work arises. Collaboration can help organisations to develop their gateways more efficiently and effectively. It can also help them to sustain the gateways in the longer term. The advantages of co-operation for organisations may include being able to:

- use established technologies, methods and practices - and avoid starting from scratch
- divide responsibilities for creating or sharing metadata records - and avoid duplication of effort
- combine effort for technical development - and avoid repetition of work and errors
- create joint publicity, training and promotion
- share staff effort (management/technical/administrative/cataloguing) - to make organisational efficiencies
- create shared strategies for long-term sustainability

All of these factors have the potential to improve the service that an organisation can offer to its target users.

For some organisations, there will be a greater imperative for collaboration if they have a remit for creating a more comprehensive service than resources will allow. This applies particularly to libraries, which are often expected to offer access to large collections, despite having limited resources to build them.

### **Disadvantages of co-operation**

There can be political or funding issues that rule out co-operation; indeed in some cases gateways will see competition as a natural alternative to collaboration! Disadvantages of gateway co-operation may include:

#### **1. Extra expense.**

To make some models for co-operation work, some extra effort will be required to set up the necessary systems. For example, to make gateways interoperable some work needs to be done on making different classification schemes, metadata formats and collection development policies compatible. In the longer term, savings may be made from having co-operative strategies but the initial setup may be too expensive to consider.

#### **2. Intellectual property rights.**

There is an issue surrounding ownership of metadata records which may stand in the way of co-operation. Gateways may have invested considerable resources into creating records and be unwilling to share them or give them away for free. The issue of intellectual property rights on the Internet is still a new one with some unresolved issues, and gateways would need to investigate these before entering co-operative agreements.

#### **3. Agreeing on aims and objectives.**

Gateways may have incompatible aims and objectives. Having developed with particular audiences in mind, they may have reservations about the value of co-operation for their users which need to be resolved. There may also be issues for funders or sponsors of gateways who have vested interests which need to be considered.

## Models for co-operation

---

In the library world, co-operative agreements that support information search and retrieval are commonplace. For example, national libraries each take responsibility for collecting materials published in their country and then offer users access to these collections via inter-library loans. Another example is the sharing of cataloguing effort, where groups of libraries work together to create union catalogues and where the catalogue records are shared and re-used by many libraries, regardless of which library actually created the record.

This co-operation enables libraries to:

- offer users access to far broader collections than could be offered by any single library
- offer users a more comprehensive catalogue than could be created by a single library
- achieve efficiencies in cataloguing and collection development without reducing the level of service to users

Such co-operation translates well into the Internet environment and the development of information gateways. Collaboration is particularly pertinent to organisations with a remit for providing access to scientific, cultural and educational resources on a large scale.

A number of different models for co-operation between gateways exist:

### **Co-operative agreements for metadata records**

Gateways can create co-operative agreements regarding metadata records:

#### **Co-operative agreements for creating metadata records**

Gateways can share the effort required to create metadata records by dividing responsibilities. For example, a group of gateways can agree that each should spend time creating records for different parts of the Internet, each focusing its efforts on records for resources in a particular subject, language or from a particular country.

#### **Co-operative agreements for using metadata records**

Metadata records can be shared and re-used, and are not confined to the service which created them or to being used in only one service. Agreements on intellectual property rights would need to be established, and work is being done in this area, but the potential exists for gateways to create agreements that enable them to offer users access to records that have been created through a distributed network of gateways. Building integrated services

Co-operation can lead to the development of integrated gateway services, which offer users access to a number of gateways via a single interface. This interface might offer different levels of functionality:

#### **Guiding users to other gateways/mirrors of gateways**

The simplest form of co-operation is for gateways to point to other gateways that might support the user group. This may involve offering a set of hyperlinks to other related gateways, or offering mirrors of related gateways where access could be improved by keeping a local copy of the service. Although each of the gateways would have to be searched serially, the user would be alerted to other gateway services which they might not have otherwise found.

#### **Fully integrating distributed gateways into a single service**

In some cases it may be easier for users if they can access many gateways simultaneously. A fully integrated service offers users the chance to select a number of gateways and then to cross-search or cross-browse all the gateways in one go. A single interface offers users a single point of access to distributed gateway services. In some cases it will not be necessary to disclose to users the fact that they are searching distributed databases.

Gateways may offer different interfaces to the same collection of metadata records. For example, a shared pool of metadata records can be developed, where each gateway contributes records to the pool, but creates its own interface to the data. In this way, different user groups can be offered a tailor-made interface and gateway service.

## Interoperability issues

---

Co-operation between gateways raises a number of interoperability issues. In the field of Internet resource discovery the term 'interoperability' refers to 'the transparent searching and retrieval of data from diverse systems and in different metadata formats' (Day, 1999).

A lot of research and development has been done on how gateways can be made to interoperate and this has highlighted the areas where standards are needed to make gateways interoperable. For gateways to co-operate they will need to work at:

- technical interoperability - search and retrieval protocols, software
- data interoperability - metadata formats, cataloguing rules

They will need to agree on:

- quality selection criteria and scope policies - to develop coherent collections and services
- areas of responsibility - to avoid duplication
- organisational/political/management issues

A fuller description of interoperability issues is given in the 'Interoperability' chapter in this handbook. However, this overview highlights some of the issues that are being tackled by existing gateways in the co-operative work described in the following sections.

### CROSS REFERENCE

[Interoperability](#)

## Practical demonstrations of co-operative work

---

Libraries and other organisations still have a lot of work to do on the political and organisational issues involved in co-operative work. However, a number of gateway projects are now able to demonstrate some of the ways in which issues of technical and data interoperability can be solved.

This section highlights a few examples of how gateways are co-operating in practical terms. These are ordered from examples of low-level co-operation, which is relatively easy to implement, to high-level co-operation, which requires agreements for a national or international strategy.

### EXAMPLE

#### **An EXAMPLE of a gateway pointing to the front pages of other gateways**

EEVL and Pinakes

EEVL (The Edinburgh Engineering Virtual Library) offers users a page of links to other high quality information gateways. This is simply a page that has hyperlinks to the front pages of other gateways; however, it may help users to find gateways which they did not know about.

- <http://www.eevl.ac.uk/>
- <http://www.hw.ac.uk/libWWW/irm/pinakes/pinakes.html>

#### **An EXAMPLE of gateways mirroring one other's services**

SOSIG/Scout Report

The UK's SOSIG (Social Science Information Gateway) and the USA's Scout Report for the Social Sciences have a reciprocal agreement to mirror one another's services, to improve access for users on both sides of the Atlantic.

- <http://scout.cs.wisc.edu/addserv/mirror/sosig>

#### **EXAMPLES of cross-searching two gateways simultaneously**

SOSIG and Biz/ed

## SOSIG and Biz/ed

In the UK, two gateways (SOSIG and Biz/ed) are offering users a service where two separate databases are simultaneously cross-searched via a single interface. Users are unaware that they are in fact searching two gateways, as the results are fully integrated.

- Go to [SOSIG](#): and search for industrial psychology.

You will retrieve records from both the SOSIG and the Biz/ed databases - displayed in a single list. Both gateways use the ROADS software which enables cross-searching

## EELS and EEVL

This is an example of two gateways based in different countries being cross-searched. Both are engineering gateways - EELS is based in Sweden and EEVL in Scotland. This is a demonstration service, but illustrates the potential for cross-searching two gateways, regardless of the fact that they are geographically separated.

- <http://roads.ukoln.ac.uk/eels-eevl/>

## An EXAMPLE of gateway standards and software that support co-operative work

### CrossROADS and Interoperability

The ROADS software has been developed specifically to support the development of gateways and to ensure that those gateways are interoperable. A demonstration of how distributed gateways can be cross-searched is available from the ROADS Web site:

### CrossROADS

- <http://www.ukoln.ac.uk/metadata/roads/crossroads/>
- <http://www.ariadne.ac.uk/issue14/metadata/>  
A paper discussing interoperability issues with metadata

## EXAMPLES of plans for integrated gateway services on a national scale

### RDN - The Resource Discovery Network

In the UK, government funding is being used to create the Resource Discovery Network - a gateway service for the higher education and research sectors. RDN will offer a single interface to a number of national subject gateways. Each of the services has its own identity and interface, but the RDN will offer another level of service to users - the ability to search for resources across several hubs at the same time.

- <http://www.rdn.ac.uk/>

### DEF Project - Denmark's Electronic Research Library

Within this project, a network of Danish libraries aims to form a virtual system to make the libraries' collective information resources (digital and traditional) available to users everywhere in the country in a simple, transparent way.

- <http://www.deflink.dk/english/def.ihtml>

## An EXAMPLE of plans for an integrated gateway service on an international scale

### REYNARD

The REYNARD project proposal suggests that national libraries in Europe should each assume responsibility for creating metadata records that describe high-quality Internet resources created in their own country. An integrated broker service will then be set up to enable each of the gateways to be accessed from a single interface and to allow users to cross-search the gateways.

- <http://www.renardus.org>

### **Key initiatives in gateway co-operation to date**

---

Are there any important initiatives in gateway co-operation? There is still much potential for co-operative strategies to be developed, particularly within the library community, but some strategies for co-operation are already developing.

#### **EXAMPLE**

##### **ROADS**

An ideal solution for co-operation would be to have agreed standards that could facilitate interoperability. The ROADS project was developed with this aim; it has created a system of software and standards for developing information gateways that have the potential to be cross-searched with any other ROADS gateway. ROADS has produced an extensive collection of software, metadata templates and guidelines, all of which are freely available.

ROADS was initially funded by the UK's Electronic Libraries Programme. The project ended in July 1999; however, ROADS continues as an open source software project, where the gateway community works collaboratively to develop the software. The ROADS community has a number of committed partners from many countries, and the software is likely to go from strength to strength.

- <http://www.ilrt.bris.ac.uk/roads/>
- <http://roads.opensource.ac.uk/>

##### **ISAAC**

ISAAC is a research project of the Internet Scout project in the USA. It aims to create an architecture that enables distributed repositories of metadata records to be cross-searched.

- <http://scout.cs.wisc.edu/research/index.html>

##### **iMesh Toolkit**

The National Science Foundation in the USA and the JISC in the UK are funding a new project (starting 1999) that will develop an architecture toolkit for distributed subject gateways. This will build on work being done within ROADS and ISAAC.

- <http://www.desire.org/html/subjectgateways/community/imesh/>

##### **DESIRE**

The DESIRE project has been funded by the European Union to develop tools and methods for organisations interested in setting up large-scale information gateways that can support European researchers. The DESIRE Web site offers information, advice and resources for gateways to use.

- <http://www.desire.org/>

##### **IMesh**

IMesh is an informal and independent group set up to facilitate international collaboration on Internet subject gateways. It was formed in 1998 after a meeting attended by staff from a number of gateways. The Web site points to a discussion forum for gateways interested in co-operation.

- <http://www.desire.org/html/subjectgateways/community/imesh/>



## Recommendations

---

Libraries, research organisations and educational establishments which are investing in the development of large-scale information gateways would be well advised to work together to create a co-operative strategy. Together they could provide the resources and expertise required to build a comprehensive collection of metadata records which describes large numbers of the high quality resources available on the Internet. Integrated services could offer users access to resources from many countries, on many subjects and in many languages.

An integrated service could offer users a valuable alternative to other Internet search tools such as search engines and directories, which are often either indiscriminate, pointing to resources of unknown quality, or popular, pointing to resources that are recreational as opposed to educational. An international network of information gateways could form the Internet equivalent of an academic research and education library, where users could go to locate high quality resources with confidence. This vision relies on co-operation and we hope that libraries and educational organisations will rise to the challenge.

## Glossary

---

**cross-browsing** - Browsing, where the Web pages contain resources from more than one gateway

**cross-searching** - Searching, where the search takes place across more than one gateway

**DEF** - Danmarks Elektroniske Forskningsbibliotek (Denmark's Electronic Research Library)

**DESIRE** - Project funded under the European Union's Telematics for research Programme to enhance and facilitate Web usage among researchers in Europe (producer of this handbook)

**EELS** - Engineering Electronic Library, Sweden

**EEVL** - Edinburgh Engineering Virtual Library

**IMesh** - An informal group for the discussion of international collaboration on Internet subject gateways

**ISAAC** - Project Isaac - A Distributed Architecture for Resource Discovery Using Metadata - managed by the Scout Project

**RDN** - Resource Discovery Network - the UK's centre for its national subject gateways

**REYNARD** - A project proposal for building a broker service to national gateways in Europe, managed by Koninklijke Bibliotheek, National Library of the Netherlands

**ROADS** - ROADS is a set of software tools to enable the set up and maintenance of Web based subject gateways.

**SOSIG** - The Social Science Information Gateway

## References

---

Biz/ed, <http://www.bized.ac.uk>

CrossROADS, <http://www.ukoln.ac.uk/metadata/roads/crossroads/>

DEF Project, <http://www.deflink.dk/english/def.ihtml>

DESIRE, <http://www.desire.org/>

EELS, <http://www.ub.lu.se/eel/>

EEVL, <http://www.eevl.ac.uk/>

IMesh, <http://www.desire.org/html/subjectgateways/community/imesh/>

ISAAC, <http://scout.cs.wisc.edu/research/index.html>

Pinakes, <http://www.hw.ac.uk/libWWW/irn/pinakes/pinakes.html>

RDN, <http://www.rdn.ac.uk/>

ROADS, <http://www.ilt.bris.ac.uk/roads/>

Scout Report Signpost, <http://www.signpost.org/signpost/>

SOSIG, <http://www.sosig.ac.uk/>

R. Heery, A. Powell & M. Day, *CrossROADS and Interoperability, Ariadne, issue 14*  
<http://www.ariadne.ac.uk/issue14/metadata/>

M. Day, *ROADS Interoperability guidelines (1999)*  
<http://www.ukoln.ac.uk/metadata/roads/interoperability/>

## Credits

---

Chapter author: [Emma Place](#)

With contributions from: Traugott Koch and Ann-Sofie Zettergren


## Section 3 : Technical Issues (Print Version)

### Target audience

---

Section 3 of this handbook is aimed at gateway staff responsible for technical implementation - Internet specialists who will manage the hardware and software and implement new technical features.

It aims to cover the important decisions that need to be made when setting up a new gateway (such as setting up the system and implementing the user interface) but also covers issues that arise in the day-to-day running of an existing gateway (such as running a link checker).

Each chapter offers some background, practical tips and hints, key references, a glossary, case studies and examples. Watch out for the  **CROSS REFERENCE** that will take you to related sections elsewhere in the handbook.

## Contents

---

Section 1 : [Strategic Issues](#)

Section 2 : [Information Issues managers](#)

Section 3 : Technical Issues

1. [System requirements specifics, hardware and software](#)
2. [User interface implementation](#)
3. [Accessibility and usability](#)
4. [Harvesting, indexing and automated metadata collection](#)
5. [User profiles](#)
6. [Interoperability](#)
7. [Scalability](#)
8. [Future proofing](#)

## 3.1. System requirements specifics, hardware and software

### In this chapter...

---

- machine and network requirements for running a gateway
- hardware and software requirements
- related technical information

### Introduction

---

This chapter provides detailed information about the hardware and software that you would need in order to set up and run an Information Gateway using the ROADS and/or Combine software.

## Background

---

The Systems Requirements Overview chapter gives an introduction to the systems-related issues which managers need to consider when setting up and running an information gateway. This chapter provides more detailed technical information about the specific software and hardware requirements that you will need to meet. It does not consider all the issues raised in that chapter. You are referred to any good UNIX systems administration book for areas not covered in detail here, since security, performance, backing up data and so on are all issues that are relevant to running any network service!

### CROSS REFERENCE

[System requirements overview](#)

## Software and hardware requirements

---

### General requirements

In order to run an information gateway you will need:

1. A machine - a computer running a UNIX-based operating system. Examples are a Sun SPARC machine running Solaris (version 2.5 or higher) or an Intel machine (typical desktop PC) running Linux. A popular information gateway will be accessed concurrently by a large number of end-users, each of whom may be searching the database. This means that it is probably worth spending money on ensuring that you have enough memory. While it is difficult to be definitive about this, because memory requirements will be specific to the operating system and software, you should probably expect to operate with a minimum of 128 Mb memory for any reasonably sized gateway. If you are considering using a PC, then it is a good idea to get the highest specification you can afford.
2. Some disk space - enough disk space to install your operating system, gateway software and Web server software and to hold your database of resource descriptions and any associated index. Assume that you'll need a gigabyte of disk space. You almost certainly won't - but in any case you probably won't be able to buy a machine with less disk space than that anyway!
3. A network connection - an Internet connection. The connection needs to be permanent (obviously!) and to provide enough bandwidth to cope with your expected number of end-users. Again, it is very difficult to be specific about this.

Don't forget about issues such as software and hardware support (and the fact that they may cost money) and think about what you are going to do when something breaks. Think about backing up your software, configuration and data. You may need a local tape drive for this or, if your organisation supports it, there may be a centralised archiving facility which you can take advantage of.

### ROADS requirements

On top of the general requirements listed above, the current release of the ROADS software (version 2) requires:

- Perl 5.002 or above (5.004 or the latest stable version of Perl 5 is recommended)
- an HTTP daemon which supports the Common Gateway Interface (CGI) specification, for example the Apache Web server. It is recommended that you use Apache, as ROADS version 3 takes advantage of mod-perl to improve its search performance

In order to run the link checking tool and its associated report generator, you will need 'libwww-perl-5', which may be obtained from [CPAN](#).



### REMEMBER

- In theory, most of ROADS can be made to run under the Microsoft NT operating system (using the GNU-Win32 toolkit from Cygnus). However, this may not be straightforward to get working and some ROADS facilities may simply never work under NT. Furthermore, there is little experience in the ROADS 'community' of using NT. For these reasons it is not recommended.

**EXAMPLE****Case study - [SOSIG](#)**

SOSIG, the Social Science Information Gateway, is a ROADS database of over 5500 Internet resource descriptions operated by ILRT at the University of Bristol in the UK. The service is hosted on a Sun Ultra-1 with 320 Mb memory running the Solaris 2.5.1 operating system. (Note that this machine also provides other services). The SOSIG data takes approximately 100 Mb of disk space and the software and gateway-specific code take up a further 50 Mb; all this data is archived across the network to a central university backup system. The service handles approximately 25,000 searches per month.

**Note:** The Web server logs associated with SOSIG are considerably larger than the data mentioned above. Depending on how much data a gateway wants to hold in its Web server access log, the disk space needed could easily be doubled (SOSIG holds approximately 400 Mb of server access logs). This kind of data will grow as the popularity of the gateway grows.

**Combine requirements**

For the [Combine](#) software, you will need:

- Perl version 5.003 or higher
- the MD5 package (from CPAN)
- the GNU 'gcc' compiler version 2.7.x or higher, complete with g++ front end and C++ libraries
- the Berkeley DB system (fetch and install the latest stable version from Sleepy-Cat Software)
- a decent version of 'make', preferably GNU's
- an HTTP daemon which supports the Common Gateway Interface (CGI) specification, for example the Apache Web server

These are in addition to the general requirements listed above.

**EXAMPLE****Case study - [All Engineering](#)**

All Engineering is a robot-generated index enabling full-text searches of all engineering pages on the Internet. The service is based on the Combine software. Holding entries for over 250,000 Web pages, the database is hosted on a Sun Ultra/Enterprise 450 running Solaris 2.6 and uses a total of 2.5 Gb of disk space.

**Glossary**


---

**CGI** - Common Gateway Interface - A standard for running external programs from a World-Wide Web HTTP server. CGI specifies how to pass arguments to the executing program as part of the HTTP request. It also defines a set of environment variables. Commonly, the program will generate some HTML which will be passed back to the browser but it can also request URL redirection. (definition from The Free On-line Dictionary of Computing)

**CPAN** - Comprehensive Perl Archive Network

**DB** - database

**GNU** - The Free Software Foundation's project to provide a freely distributable replacement for Unix.

**ILRT** - Institute for Learning and Research Technology

**ROADS** - Resource Organisation and Discovery in Subject-based services - a set of software tools to enable the set up and maintenance of Web based subject gateways.

## References

---

- All Engineering, <http://www.lub.lu.se/eel/ae/>
- Apache, <http://www.apache.org/>
- BerkeleyDB, <http://www.sleepycat.com/>
- Combine, <http://www.lub.lu.se/combine/>
- CPAN, <http://www.sn.no/libwww-perl/>
- Cygnus, <http://www.Cygnus.com/>
- GNU, <http://www.gnu.org/>
- Linux, <http://www.linux.org/>
- Perl, <http://www.perl.com/>
- ROADS, <http://www.ilrt.bris.ac.uk/roads/>
- Sleepy-Cat Software, <http://www.sleepycat.com/>
- SOSIG, <http://www.sosig.ac.uk/>
- AE. Frisch, *Essential System Administration (2nd ed.)* (ISBN: 1-56592-127-5)
- M. Loukides, *System Performance Tuning* (ISBN: 0-937175-60-9)

## Credits

---

- Chapter author : [Andy Powell](#)
- With contributions from: Paul Hollands

## 3.2. User interface implementation

### In this chapter...

---

- general Web design issues: look 'n' feel, frames or no frames?
- design implementation issues specific to information gateways
- informing the user about the gateway
- the search interface and the browse interface
- combining searching and browsing (including cross-searching and cross-browsing)
- the thesaurus interface
- the cataloguing interface

## Introduction

---

The chapter entitled User Interface Design introduced the major issues in the design of Web interfaces and in the collection of data to help inform a user interface design specification. The present chapter will look in more detail at those issues which are particularly relevant to the design of information gateways. Although some of the answers to the questions discussed here will be determined by your choice of software for running your gateway, the following points should still be considered before committing your institution to a particular solution.

### CROSS REFERENCE

[User interface design](#)

## Background and Overview

---

The 'user interface design' chapter reviews the reasons why good interface design is necessary. However, there are important issues to consider which result from the limitations of the Web and HTML as a presentation tool and formatting language respectively, as well as from inconsistencies in the capabilities of different clients and the machines they run on. Both of these factors can cause problems in the attempt to realise your design.

Problems of the first sort can usually be solved with a little ingenuity on the part of the Web designer, together with the use of helper technologies such as server-side scripting and stylesheets. The second type of problem is related to accessibility and usability issues and is covered in the chapter 'Accessibility and usability'.

### CROSS REFERENCE

[Accessibility and usability](#)

This chapter will therefore describe the approaches to implementing information gateway design that have been found to be of practical value within the gateways produced as a result of the work of the DESIRE projects, together with the results of their continuing experimental development.

## Recommendations

---

### General Web design issues

Many of the issues relating to good design practice for Information Gateways are common to all Web sites and have been covered in the User Interface Design chapter.

### CROSS REFERENCE

[User interface design](#)

### Look 'n' feel

The look of the site as a whole is best managed with mechanisms that allow for easy global control of style and content. [Cascading Style Sheets](#) (CSS) are an obvious choice, although care should be taken to test these against a variety of browsers and browser versions; there is still some incompatibility between Netscape Navigator and Internet Explorer and style sheets will not work on early versions of either. It is consequently vital to check your site on a number of different browsers to see how much your style sheets degrade on earlier versions. A useful online resource describing differences between various browser CSS implementation bugs is '[CSS Bugs and Workarounds](#)'

An additional mechanism for adding common elements to the site's pages is the use of Server-Side Includes (SSIs). These provide an excellent way to add components such as navigation bars (or style sheet references), as well as other common features such as feedback links and site logos, to sets of pages within the site. They work by using special tags which can be added to the HTML of a page and which cause the server to insert standard content at those locations. However, since the server needs to parse each of these pages before sending them on to a client, SSIs will reduce server performance.

Both of these methods can also be applied to the display of search results, which will consist of pages generated on the fly (see the section 'Presenting search results').

### Frames or no frames?

There is some controversy over whether frames should be used in Web sites (e.g. '[Why frames suck most of the time](#)'). As a means of enhancing navigation about a site, they can be very effective if used carefully; for instance a single frame down one edge could contain links to the various sections of the site. They can also make it easy for the user to return to your site having selected a link from their search results, since the remote site can be displayed within a frame.

However, the navigation mechanisms can be provided as easily with SSIs; and the frames technique is generally frowned upon due to the problems of bookmarking, the copyright issues that arise from displaying a remote site within your own, and the reduction in screen space that results. There is also the potential problem of 'frames within frames' if the remote site also uses them.

## Design implementation issues specific to information gateways

---

Apart from general Web site design considerations, a number of interface issues need to be addressed which relate specifically to the nature of an information gateway. The main challenges involved are those of informing users what information the gateway contains and of enabling users to search that information sufficiently well to obtain the results they require. A third consideration concerns the manner in which search results are displayed to the user.

It should be borne in mind that many users are not expert at searching databases and may not even be very familiar with the structure of the subject covered by the gateway. These are problems which have been faced by information professionals ever since the introduction of end-user searching with the development of CD-ROM databases.

This section will look at these specialised user interface design issues.

### Informing the users about the gateway

Our user studies have shown that most gateway users do not understand the difference between information gateways, directory services such as Yahoo! or search engines such as Alta Vista. It is also clear that few users make use of any search engine's full functionality. It is therefore important to provide sufficient text to explain what the gateway consists of and how it works, including its aims and policies, whilst accepting that most users do not like reading much text from the screen and that they should be presented with an uncluttered and simple looking interface which will not intimidate them.

The usual attempt to solve this apparently impossible task is to provide information in the form of 'help' files but these are also unlikely to be read by the majority of users without some encouragement. Methods which may have more success include:

- context-sensitive help, where a 'help' link or icon will give information relevant to the page being viewed
- FAQs, which list the questions that users have been found to ask most often
- tips, which may be displayed randomly on a search page or which can appear with advice under certain conditions, for instance when a user is getting no hits

The search pages of the [Social Science Information Gateway](#) (SOSIG) and of [OMNI](#) demonstrate different methods of linking to 'help' information.

### The search interface

Here also the main problem with presenting an interface to a search engine lies in making the full functionality of the engine available to the user in such a way that they can understand and use all its features without being intimidated. The usual approach is to provide two interfaces: one for simple searching and one containing the more advanced features.

The search functionality available will obviously depend on the database and application software chosen to run on the catalogue, but advanced features will usually include options such as Boolean searching (may be implemented as all or any of terms in the query), phrase searching, searching by field (title, keyword, resource type, date range, etc.), case-sensitive searching and various methods of truncation or stemming. The usual way for the user to send in their search terms and option choices is by means of a typical HTML form. The selection of choices may be made with any of the standard HTML form options: radio buttons, checkboxes or pull-down menus. A common way of providing a 'simple' search interface is to provide default values of these options as 'hidden' values in the HTML form code.

Unfortunately, experience from general Web search engines (e.g. <http://www.useit.com/alertbox/9707b.html> and <http://www.useit.com/alertbox/9707b.html>) and information gateways shows that advanced features are seldom used; for example, SOSIG has under 10% of its searches made from its advanced search page. This may be because users fail to understand their usefulness or are simply put off by a link that says 'Advanced search'. Help features, as described above, can ease this problem, but the interface designer should be aware of this issue when designing any 'advanced' search page.

See the SOSIG [advanced search](#) page.

## Presenting Search Results

It is useful to provide users with the alternatives of displaying results by title alone or giving the full description, possibly including other fields such as keywords. A third option might be to display the full set of metadata contained in the record.

With 'titles only' selected, the full set of results can be displayed; when displaying full record details it is necessary to limit the length of the pages produced, otherwise the files transmitted can be very large, take too long to download, and require the user to do too much scrolling. Two methods of achieving this are by placing a limit on the number of results that will be displayed, requiring the user to further refine their search, or by displaying results on a number of separate pages.

### EXAMPLE

#### Biz/ed search result views

Biz/ed uses the functionality of the ROADS software to offer the option of returning search results as either titles only or as full records. The user is free to choose which option they prefer:

- [Biz/ed search page](#)
- [Biz/ed titles only search result \(search term = Marx\)](#)
- [Biz/ed full record search results \(search term = Marx\)](#)

With sets of data containing a few thousand records, the former method is quite practical, but becomes less so as the number of records in the database increases resulting in a corresponding increase in the average number of hits produced by a search. The average number of hits produced should therefore be monitored and the limit adjusted accordingly so that the server refuses only a small proportion of searches. Any such refusal to transmit too large a results set should be combined with mechanisms for narrowing the search, perhaps with a link to the advanced search page or to a thesaurus (see below). Alternatively, only the first portion of the results could be displayed, provided that some sort of ranking mechanism were being used to ensure that the most relevant results were shown (see below).

The other option is to divide the results set over several pages. Whether results can be transmitted in this manner will depend on the search application used (for example, Z39.50 permits this, but Whois++ does not). A ranking mechanism is also useful with this method.

It is usual to rank the results of keyword searches to ensure that the most relevant records come at the top of the list. This is usually accomplished with an algorithm which looks at the frequency with which search words appear in the records, with weightings applied depending on the location of the term (e.g. terms in the title, first paragraph and metadata fields will have a high weighting factor). It may be possible to amend or replace an existing ranking algorithm, perhaps by adjusting the weightings or by introducing factors based on user preferences (such as educational level of material or resource type), depending on what information is available in the records.

You might also consider including a few easy to implement but very useful things in your search results pages:

1. Repeat the original search query prominently on the results page. As users browse through search results, they may forget what they searched for in the first place. Remind them. Also include the query in the page's title; this will make it easier for users to find it in their browser's history list.
2. Let the user know how many matches to their query have been retrieved. Users want to know how many documents have been retrieved before they begin reviewing the results. Let them know; if the number is too large, they should have the option of refining their search.
3. Let the user know where he or she is in the current retrieval set.
4. Always make it easy for the user to revise a search or start a new one. Give them these options on every results page, and display the current search query on the 'Revise Search' page so they can modify it without re-entering it.

(after Rosenfeld and Morville, 1998, p. 121)

## Browsing the catalogue

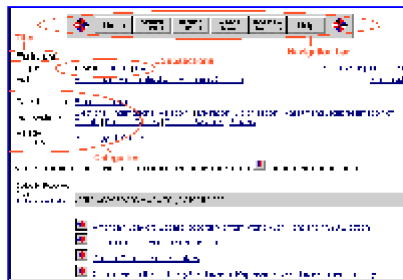


The majority of information gateways provide browsing access to their collections as well as keyword searching. This is achieved by manually (or automatically) classifying individual resources according to a hierarchical classification scheme. Records for resources with the same class number (they may have more than one each) are displayed on the same page, with pages structured according to the classification scheme hierarchy. It is not usual to display the class numbers themselves, since these are of little interest to users, but to display only the title of the section.

### CROSS REFERENCE

#### [Subject indexing and classification](#)

There will need to be hypertext links between the different sections of the classification scheme structure, including links to parent, child and possibly 'related' sections. Simple HTML hypertext links can be used to represent the structure of the scheme, but it is important that the design enables easy navigation without the user's getting lost.



*[Politics browse section from SOSIG](#)*

Depending on the facilities offered by the application software, the browse pages may be generated on the fly or periodically generated with a script; the latter method is used by the ROADS software. The script that generates the page will in many cases simply list the resources in alphabetical order but can also be used to group or filter them according to some other criterion such as resource type or country of origin. With a periodically generated set of pages, these latter options can be implemented simply by producing separate pages for each possible view.

To enable the records to be split up into the different browse sections, a search using a class number field is made, or else the records themselves can be stored in directories whose hierarchical structure corresponds to that of the classification system.

### **Combining searching and browsing**

Browsing and searching can also be combined to allow a simple search to be made from within the browse pages. This facility may offer the option of searching only those resources listed within the currently viewed classification section and all child sections, rather than the database as a whole.

One method of accomplishing such a search is to hold the records in a file system whose hierarchical structure mirrors that of the classification scheme and restrict the records searched to those within the current directory plus child directories.

An alternative approach is to perform a keyword search for the class numbers themselves in addition to the user's search terms. This can be problematic, however, as the search can end up involving a large number of child sections, requiring a complicated Boolean OR search that inevitably slows down the search engine. This problem may be overcome if the class numbers permit meaningful truncation or, if the notation of the classification system is not constructed in this manner, an alternative, hidden representation of the class numbers could be devised for the purpose which did permit it.

### **Cross-searching and cross-browsing issues**

Methods of enabling the cross-searching and cross-browsing of Information Gateways are given in the chapter on Interoperability. However, there are a number of issues concerning the way that cross-searching and cross-browsing are presented to the user.

Firstly, there is the question of whether a cross-searching facility should be made obvious to the

user or kept hidden. If the mechanism is made open, how should it be presented to the user in a way they can understand? It would certainly be useful to provide information on each gateway concerning scope and selection criteria and a mechanism for selecting which gateways will be searched.

With cross-browsing, there is also the question of what is actually meant by the term. One approach (used by the [Social Science Information Gateway](#)) is to enrich the holdings of one catalogue with links to the records of one or more other catalogues, the links being placed in the browsing structure alongside references to local records. An alternative approach to cross-browsing is simply to insert links within each browse section to the equivalent sections of other gateways. The user is then actually browsing across catalogues.



These areas are currently being worked on within the Desire project and research findings will be published in the near future.

- [Desire research findings](#)
- [Desire project deliverables](#)

A further issue connected with the presentation of results of cross-browsing and searching concerns how or whether individual records should be differentiated by their origin. This could be done with additional text or copyright declarations or by the use of different icons. But this may be considered unnecessary (as far as the user is concerned, though perhaps necessary because of intellectual property rights considerations) and potentially confusing.

A discussion of how cross-browsing may be achieved is given in the Interoperability chapter.



[Interoperability](#)

#### EXAMPLE

##### **Cross-searching results interface**

For an example of the results of a search across the catalogues of the [Social Science Information Gateway](#) (SOSIG) and [Biz/ed](#):

Search for [banking AND Europe](#)

For an example of a browse section within SOSIG that actually contains records from the Biz/ed catalogue: [SOSIG economics section](#)

##### **The thesaurus interface**

The Subject indexing and classification chapter discusses the issues involved in choosing a thesaurus for enhancing searching. In most cases an existing thesaurus relevant to the subject coverage of your information gateway will have been chosen and a local copy obtained (subject to agreements with the copyright holder).



[Subject indexing and classification](#)

To ensure that terms selected from the thesaurus produce useful results from your catalogue, we recommend that the local copy be a subset of the full thesaurus, which includes only those terms used in your catalogue. This can be accomplished by periodically running a script which compares the thesaurus terms against the catalogue's index. A decision will have to be taken as to whether the controlled terms from the thesaurus will be searched against all text in the catalogue records or restricted to terms in a keyword field.

It is likely that the software for the local copy of the thesaurus will have to be created in-house. It should allow easy navigation through the hierarchy of terms and ideally allow searches of the

catalogue to be performed automatically from those terms selected by the user.

#### EXAMPLE

##### Example of a gateway using a thesaurus

SOSIG uses HASSET (Humanities And Social Science Electronic Thesaurus), created by The Data Archive in the UK. SOSIG cataloguers use HASSET to generate keywords. The thesaurus offered to SOSIG users however, is a customised version, containing terms which appear both in HASSET and the SOSIG index, enabling users to search the SOSIG catalogue using the HASSET interface.

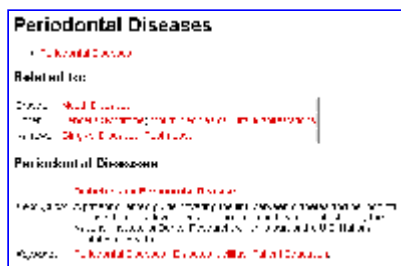
- [HASSET](#)
- [SOSIG Thesaurus](#)

A useful feature to add is the option of searching for the selected term together with all 'child' terms - a feature often known as an 'explode' option. As with searching by keyword within the browse sections of the catalogue, this can involve a complicated Boolean OR search, which is unacceptably slow. Similar techniques to those described in the section on combined searching and browsing could possibly be used to remedy this; for instance, by using an alternative representation of the keywords which could be used with truncation. As with the catalogue itself, it will usually be possible to browse through the hierarchical structure of the thesaurus as well as to search it by keyword. There may also be an alphabetical index of terms with links to the thesaurus. Browsing the thesaurus can be accomplished with hypertext links between related terms, with parent, child, related and non-preferred terms listed with the currently selected term.

An alternative way to use the thesaurus for access to catalogue records is to produce a list of all records that contain the currently selected term. This turns the thesaurus into an alternative classification system.

It is quite common for users to become confused and to believe they are actually searching the catalogue rather than the thesaurus; hence it is necessary to ensure that the thesaurus has a very different look and feel from the catalogue itself.

See the example from OMNI below for an illustration of this.



*[MESH subject heading from OMNI](#)*

#### The cataloguing interface

All the interface implementation issues discussed so far concern the users of the catalogue. However, you also need to consider the way in which the cataloguing interface is implemented in order to ensure efficient data entry by the cataloguers of the system.

#### CROSS REFERENCE

##### [Cataloguing](#)

As with many other implementation issues, the cataloguing interface will depend largely on the application being used. The following features should be considered when deciding on a system or designing one in-house:

- the ability to locate any record quickly and bring it to an editing screen
- the facility to perform global edits

- a set of authority lists for adding class numbers, controlled vocabulary terms (possibly via access to the thesaurus), and any other data that needs to be in a standard format, such as country codes, language codes, etc.
- a variety of standard templates if different formats are used for different types of resource
- the ability to store completed records for proof checking before they are entered into the catalogue
- help facilities

## Glossary

---

**Boolean searching** - The use of the "Boolean operators" (AND, OR, NOT) in keyword searching to combine keywords and so control the resulting matches and make more precise searches.

**Cascading Style Sheet (CSS)** - A style sheet language that allows the authors of Web pages to separate the content of HTML files from form and appearance. Style sheets enable Web authors to apply a uniform style to a group of documents in a web site.

**Cross-browsing** - Browsing, where the Web pages contain resources from more than one gateway

**Cross-searching** - Searching, where the search takes place across more than one gateway

**DESIRE** - Project funded under the European Union's Telematics for research Programme to enhance and facilitate Web usage among researchers in Europe (producer of this handbook)

**HASSET** - Humanities And Social Science Electronic Thesaurus, produced by The Data Archive in the UK

**MESH** - Medical Subject Headings

**OMNI** - Organising Medical Networked Information (UK national gateway)

**Server-Side Include (SSI)** - The facility provided by several HTTP servers, e.g. NCSA httpd, to replace certain HTML tags in one HTML file with the contents of another file at the time when the file is sent out by the server, i.e. an HTML macro. Definition taken from [NCSA httpd tutorial](#)

**SOSIG** - The Social Science Information Gateway

**Template** - A form based on a metadata format with fields for the key attributes required to describe a resource and space to add values for each of these attributes to create a catalogue record.

**Thesaurus** - A thesaurus represents a collection of organised knowledge, often based on an abstract classification scheme, which provides a "map" of some subject domain. It is used by professional indexers as a source of controlled language (Centre for Interactive Systems Research definition)

**Whois++** - An Internet directory services protocol

**Z39.50** - A NISO standard for an applications layer protocol for information retrieval which is specifically designed to aid retrieval from distributed servers.

## References

---

Biz/ed, <http://www.bized.ac.uk/>

CSS Bugs and Workarounds, <http://css.nu/pointers/bugs.html>

HASSET, <http://dasun1.essex.ac.uk/services/zhasset.html>

OMNI, <http://www.omni.ac.uk/>

SOSIG, <http://www.sosig.ac.uk/>

W3C Cascading Style Sheets, <http://www.w3.org/Style/css/>

L. Rosenfeld & P. Morville, *Information Architecture for the World Wide Web* (O'Reilly, 1998).

Jakob Nielsen 'Why frames suck most of the time'  
<http://www.useit.com/alertbox/9612.html>

## Credits

---

Chapter author: [Phil Cross](#), [Martin Belcher](#)

With contributions from: Jan Chipchase

## 3.3. Accessibility and usability

### In this chapter...

---

- drawing up accessibility guidelines for your gateway
- implementation of accessibility guidelines
- validating your gateway's accessibility

### Introduction

---

The issues of good accessibility and usability are closely linked. Their importance has been emphasised in previous chapters of the handbook. How can these issues be best tackled and implemented in the development of a new gateway or the modification of an existing one?

#### CROSS REFERENCE

[User interface design](#)

### Accessibility and usability for your gateway

---

The accessibility and usability criteria of your gateway should have been drawn up after some degree of user consultation. Ideally, the user consultation will have produced a user interface design specification; The specification should contain particular information such as the gateway name, section division naming (if appropriate), structure and information architecture. Guidelines or parameters such as maximum page size (pixels and/or bytes), maximum download times, colour palette size and makeup, colour scheme and use of images will also form part of the specification. An ideal end result might be a document in the form of a checklist, against which a design can be developed and checked.

Remember that a checklist which contains too many items can be unusable in itself. Test a prototype version of your checklist to see if it is usable, before rolling it out to all developers. A design specification will probably be divided into several areas.

#### Usability issues

What usability issues will the gateway conform to? Guidelines here might be:

- users will be able to search from every page
- users will be able to search with one click
- help (or perhaps context-sensitive help) will be available within every page
- users will never be more than one click away from the homepage
- feedback and comment facilities will always be no more than one click away
- feedback will be provided to users on their actions (e.g. holding page while the search is being processed, guidelines on using search result pages)
- users will not need to understand technical terminology (i.e. Boolean search rules) to carry out basic tasks

#### Site structure and navigation

It seems obvious, but some of the key problems with Web sites arise from the naming of sub-sections and the associated navigation of them. Fortunately, information gateways have common key sections which can easily be worked into a navigation system and which are almost universally understood (subject-specific and specialised gateways may differ in this area and so may be tailored to the user community). Section names often include:

- home
- search
- browse

- help
- what's new

### Accessibility issues

What accessibility criteria will the gateway conform to? Fortunately, a definitive set of accessibility guidelines already exists in the form of a [W3C Recommendation: Web Content Accessibility Guidelines 1.0](#). It would save time and effort to adopt some or all of these official guidelines. The exact guidelines that are used may vary from gateway to gateway, as there are many recommendations and it may not be realistic to implement them all. Luckily, the guidelines have been prioritised in a way that makes it easy to see which accessibility issues have the greatest influence on potential users:

- priority 1: must do
- priority 2: should do

(see '[Disabled Accessibility: The Pragmatic Approach](#)')

You might decide only to use items in the 'Priority 1' checklist and a selection of those from the lower priority groups, for example:

1. Provide a text equivalent for every non-text element (e.g. via 'alt', 'longdesc', or in element content). Non-text elements include: images, graphical representations of text (including symbols), image map regions, animations (e.g. animated GIFs), applets and programmatic objects, ASCII art, frames, scripts, images used as list bullets, spacers, graphical buttons, sounds (played with or without user interaction), stand-alone audio files, audio tracks of video, and video.
2. Ensure that all information conveyed with colour is also available without colour, for example from context or markup.
3. Clearly identify changes in the natural language of a document's text and any text equivalents (e.g. captions).
4. Organise documents so they may be read without style sheets. For example, when an HTML document is rendered without associated style sheets, it must still be possible to read the document.

### Implementing accessibility guidelines

The simplest way to implement and check that your gateway meets its accessibility and usability requirements is to use a simple 'checklist' during development of the interface. Developing the user interface as a series of templates, separated from the technology of the gateway, makes changing aspects of the interface much easier. As the interface develops it can be continually checked against the checklist of requirements.

When a gateway's interface is complete, it is often worth stating that the site conforms to certain guidelines (e.g. HTML 4.0, Bobby Approved, Web interoperability); however, do not do this on your most commonly accessed pages (e.g. the home page or the search page) but rather confine this information to an 'about' section or page.

### Validating your gateway's accessibility



#### Accessibility validating using Bobby

Bobby is a Web-based tool which analyses Web pages for their accessibility to people with disabilities. Bobby's analysis of accessibility is based on the World Wide Web Consortium's (W3C) Web Content Accessibility Guidelines.

Bobby also analyses Web pages for compatibility with various browsers. Analysis is based on documentation from browser vendors, when this is available. Bobby automatically checks sites for compatibility with HTML 4.0. For accessibility and tag compatibility with browser specifications other than HTML 4.0, use the Advanced Options. Once your web site receives a Bobby Approved rating, you are entitled to use a Bobby Approved icon on your site.

Bobby is available as a free downloadable application which allows you to check

Bobby is available as a free downloadable application which allows you to check multiple local files or entire Web sites in one operation. The application runs the same page-checking code as the online version. Bobby is a very useful resource which should be used by all gateway developers and maintainers.

- Bobby: <http://www.cast.org/bobby/>

## Usability into the future

---

It is worth noting that Web-related technologies change, users change and information changes. However, seldom do any of these variables change at the same time. The result is that you should always be aware that the criteria for usability and accessibility are not set in stone. Along with other aspects of the gateway, these criteria should be reviewed from time to time and, if need be, adjusted to meet changes and developments. It should be noted that users rarely change as quickly as everything else around them! Caution is therefore advisable when implementing any user-side technological changes.



- Adopting a Web accessibility policy makes your Web site more usable for all users.

## Glossary

---

**Accessibility** - the characteristics of Web content and whether or not it is accessible to people with disabilities

**Usability** - the degree of ease with which human beings can interact with an object, in particular a computer system

**W3C** - World Wide Web Consortium

## References

---

Bobby, <http://www.cast.org/bobby/>

Disabled Accessibility: The Pragmatic Approach  
<http://www.useit.com/alertbox/990613.html>

Jacob Nielsen's Alertbox Column  
<http://www.useit.com/alertbox/>

List of Checkpoints for Web Content Accessibility Guidelines 1.0  
<http://www.w3.org/TR/WAI-WEBCONTENT/checkpoint-list.html>

L. Rosenfeld & P. Morville, *Information Architecture for the World Wide Web* (O'Reilly, 1998).

J. M. Spool et al., *Web Site Usability: A Designers Guide* (Morgan Kaufmann Publishers Inc., 1999).

W3C, *Web Content Accessibility Guidelines 1.0*  
<http://www.w3.org/TR/WAI-WEBCONTENT/>

## Credits

---

Chapter author: [Martin Belcher](#), [Phil Cross](#)

With contributions from: Jan Chipchase

## 3.4. Harvesting, indexing and automated metadata collection

### In this chapter...

---

- The technical aspects behind automatic collection of Internet resource descriptions and how to make good use of the results
- The software used by the DESIRE II project is reviewed - possibilities and limitations
- Try for yourself; set up a Harvested Information Gateway! We'll show you how to do it

### Introduction

---

This chapter provides a starting point for technical specialists who are considering using harvesting, indexing and automated metadata collection within their information gateway. An information gateway which works like this consists of three separate mechanisms:

- A robot which collects resource descriptions from the Web according to a set of rules. Care must be taken in order to assure that the robot detects and saves any metadata provided within the resource. [NetLab](#) develops and maintain a Web harvesting system called [Combine](#).
- The collected resources must be indexed and made available using a server that can process queries and requests for information retrieval. DESIRE II uses the [Zebra](#) search engine from [Indexdata](#) which implements the [ANSI/NISO Z39.50](#) search and retrieval protocol.
- Finally, the indexed resources hosted by the server must be made readily available to the end-users. Thus we need a Web interface that is able to communicate with the server, i.e. compliant with the [ANSI/NISO Z39.50](#) protocol, and which can respond to end-users' requests. There exist a few gateways with such an interface. We will use the [Europagate](#) service provided by [dtv](#).

The main software components used in the DESIRE II project are reviewed. The rest of this chapter describes how to glue the different pieces together into a running environment that can accommodate further development.

### Background

---

The core function of an information gateway is to make bibliographic records available for advanced searching. The [ANSI/NISO Z39.50](#) protocol is specially designed to support very detailed request and retrieval sessions. That is why the Desire project uses the [Zebra](#) server software which implements that very protocol. Since [ANSI/NISO Z39.50](#) isn't very widely supported (none of the major Web browsers provides a client) we need to use a gateway. The gateway's main functionality is to channel requests passed via HTTP to a Z39.50 server and return an appropriate response. It also has to keep track of all the different sessions for all users who access the gateway. Finally, we obviously should have a robot to collect the Web resources in the first place. There are many robots available, but we need one that can deal with our particular interest in metadata as well as our need to adjust robot output in a way that makes it easily available to the [Zebra](#) server. [Combine](#) fulfils both these requirements.

### Harvesting and Combine

---

The harvesting metaphor was coined because of the strong similarities between the automated collection of Web resources and real-world harvesting. Both of these tasks raise three key issues:

1. What sort of crop are we interested in and where do we find it?
2. How do we harvest?
3. Can we keep the weeds out?

The first question is concerned with how best to discover Internet resources and is primarily a matter of manual selection. Those aspects are described in a separate chapter.

#### CROSS REFERENCE

[Resource discovery](#)

It does, however, highlight an important problem that begs for computerised support. A harvester works very well on a field of corn but it performs poorly in other contexts, for instance when we're looking for rare mushrooms in a forest. We simply cannot take everything and then sift the mushrooms from the wood, grass and pebbles. A similar line of reasoning applies to a Web robot.



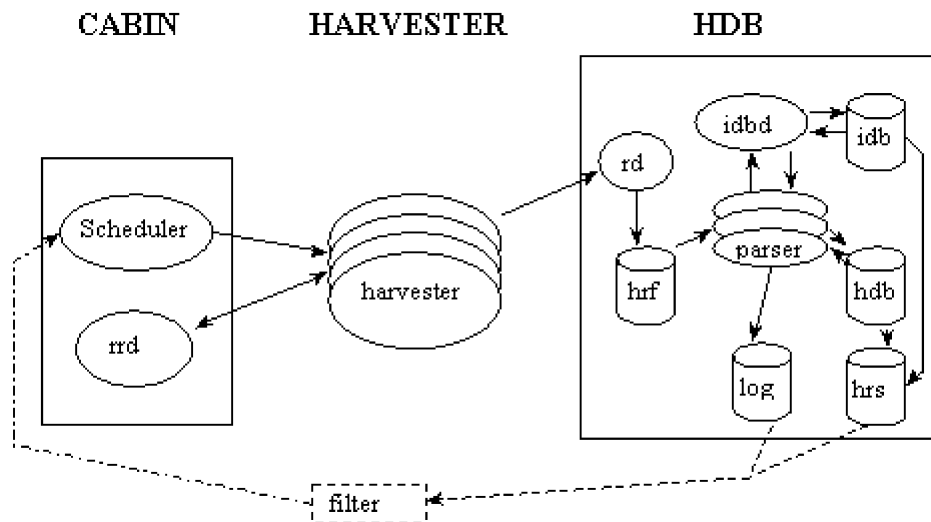
It would be a huge waste of time and resources to make a robot crawl around the entire .com domain in order to harvest any page concerning the sale of fountain pens. While it is possible to employ subject specialists to detect valuable Web resources and librarians to catalogue them, such an approach is relatively expensive. For this reason it is tempting to design a Web robot that, when given a promising starting point, is able to select which trails to follow.

### EXAMPLE

#### EELS and All Engineering

An interesting attempt to address these matters has been made within the DESIRE II project. Read about [EELS and All Engineering](#).

The last two questions are easier to approach from the point of view of an information analyst who wishes to design a Web robot so we'll dispense with the agronomics. Instead we shall turn our attention to how the Combine system is designed to serve as an integral part of an information gateway. Combine is an open, metadata-aware system for distributed, collaborative Web indexing and it is freely available. It consists of a scheduler, a couple of robots, and receivers that process and store robot output.



1. The scheduler is loaded with a set of nodes called JCFs which each contain an URL and some meta information. Depending on an internal set of rules that are configurable, the scheduler selects the next URL to be processed and launches a robot (harvester).
2. The robot visits its target server and retrieves data. It is designed to be very polite and well mannered towards the targeted server in order to keep its administrator happy. Data is delivered via a receiver (rd) and written to a depot (hrf) where the parsers can access it.
3. The parsers are able to detect metadata as well as metadata formats such as [Dublin Core](#). The parsers mark up all detected metadata and hyperlinks in accordance with a special format. Parser output is stored in a tree-like manner directly on the filesystem under the hdb directory. The hyperlinks that constitute a complete URL can be recycled, thus allowing recursive harvesting of a Web site.

You are strongly recommended to visit the Combine home page <http://www.lub.lu.se/combine> to get a general overview before trying to install and run Combine. Note that some information on the Combine home page may be a bit out of date.

#### Installing and running Combine

Before you start, make sure you have:

- a system running your favourite UNIX flavour. Combine has been successfully installed under various versions of Linux and Solaris 2.5 and higher
- Perl version 5.003 or higher, including the MD5 package
- gcc 2.7.x or higher, complete with g++ front end and C++ libraries
- the [Berkeley DB system](#); fetch and install the latest stable version from [Sleepy-Cat](#)

### Software

- a decent version of make, preferably GNU's
- created a top level directory within which everything will be built. Call it, for instance, **DESIRE2**

### Installation

1. Fetch the latest stable distribution from the [Combine home page](#).
2. Unpack the tarball; type **'tar xzvf combine-???.src.tgz'**.
3. Enter the unpacked directory, from now on referred to as **'combine-src/'**. Type **'cd combine-src/'**.
4. Edit the Makefile. Most users will only need to make three changes:
  - a) Set **'HOME\_ALL'** to indicate where to build Combine. Make sure that the directory exists. The build directory will be referred to as **'COMBINE/'**.
  - b) Set **'DB'** to the directory where your Berkeley DB system is located.
  - c) Uncomment any line concerning your OS under the platform specific section.
5. Type **'make; make install'**.
6. Everything should go smoothly but don't hesitate to use the [mailing list](#) if you have any trouble installing the Combine software.

### Configuration

1. Create a file, say, **'starturls.txt'** in your **'COMBINE/etc'** directory. Put the URLs you wish to harvest on separate lines in **'starturls.txt'**. Remember, Combine supports recursive harvesting so you don't need to provide URLs to all individual pages on a domain.
2. The Combine system's ability to recursively harvest a Web site poses a problem. We may very well want to restrict our search for Web resources to a specific host or domain or similar. To do this, edit the **'config\_allow'** and **'config\_exclude'** files in **'COMBINE/etc/'**. The files are configured by means of regular expressions similar to Perl's and they contain a few typical examples.
3. Edit the file **'COMBINE/etc/combine.conf'** and provide the necessary information.
4. Browse the **'COMBINE/etc/config\_binext'** and **'COMBINE/etc/config\_parsable'**.

### Running Combine

Note that this example is intended to show what a Combine session looks like and is therefore run by hand.

1. Type **'cd COMBINE/'** since some scripts depend on being run from that directory.
2. Type **'bin/start-cabin'**.
3. **'bin/start-hdb 2'** where '2' tells Combine that we want 2 parsers.
4. **'bin/start-harvester-local 4'**, twice as many harvesters.
5. Prepare the scheduler. Type **'bin/sd-ctrl.pl open; bin/sd-ctrl.pl pause'**.
6. We're all fired up and ready to feed Combine with input. This is done by piping our URLs in **'COMBINE/etc/starturls.txt'** through a set of filters:
  - a) The first filter **'bin/selurl.pl'** applies the rules in **'config\_allow'** and **'config\_exclude'** and it can be omitted.
  - b) 'jcf' stands for job control format and it is Combine's internal representation of an URL. Since all URLs must be formatted this way, the filter **'bin/jcf-builder-uniq.pl'** is useful.
  - c) Finally, we load our jcfs into the scheduler with **'bin/sd-load.pl'**.
7. Let's put it all together:  
**'cat start-urls.txt | bin/selurl.pl | bin/jcf-builder-uniq.pl | bin/sd-load.pl'**  
**Note:** Only **'bin/sd-load.pl'** affects the state of Combine, so don't be afraid to experiment with the others.
8. Launch Combine with **'bin/await-harvest.pl 1'**.

### Now what?

If everything went fine, there should be a file entry with a **'rec'** suffix for each harvested Web page under the **'COMBINE/hdb/'** directory. Take some time to browse the directories to see what has happened during your first Combine session. In order to harvest all interesting links that resulted from the this session, simply type:

**'bin/new-url.pl | bin/selurl.pl | bin/jcf-builder-uniq.pl | bin/sd-load.pl'**

People who are more interested in getting things done rather than wasting time with low-level

Combine details may irritably ask themselves: 'Isn't there any high-level interface available to all this nonsense?' Fortunately there is. Browse the html document [cje/cje.html](#) and find out how to install and run the Combine Job Editor. Note that you need a Web server to take full advantage of this package.

### Zebra and Z39.50

Zebra is an indexing system and a retrieval engine attached to a Z39.50 server. The following introduction to Z39.50 comes from a [document](#) at [Indexdata](#) describing Zebra.

The ANSI/NISO Z39.50-1995 standard presents a model of a very flexible, general-purpose information management and retrieval system. The intent is that this model should be placed 'in front' of new and existing information systems, to provide a uniform interface to client applications. This in turn provides the user with a number of benefits, including a uniform interface to many different kinds of information sources - hopefully tailored exactly to his specific needs by the provider of the client software. Z39.50 allows many different systems to look the same to the individual user, and it allows the individual information system to appear in many different forms, to suit the varying preferences and requirements of the users.

The quotation above should encourage you to believe that Zebra will somehow index and answer Z39.50 queries on, say, the stuff that Combine recently fetched from the Web.

### Installing and running Zebra

#### Installation

1. Get zebra and yaz from [Indexdata](#).
2. Unpack the tarballs from the **DESIRE2** directory.
3. Installation is simple. Enter each directory and type '**configure; make**'. Make sure that you build yaz before you build zebra.
4. Check your zebra/index/ directory for two executables: zebraidx and zebrasrv.

#### Configuring and running Zebra

1. Download the [configuration files](#) and unpack them with '**tar xzvf zcfg.tgz**'. Enter the new directory **zebraindex**.
2. Create a link to the data collected by Combine. Type '**ln -s ../COMBINE/hdb hdb**'.
3. Browse the configuration in zebra.zfg and check all paths. Try to create an index by typing: '**./zebra/index/zebraidx -c zebra.cfg -g index update hdb >&! index.log**'
4. Start the zebra server. Type: '**./zebra/index/zebrasrv -c zebra.cfg tcp:host.domain:1101 &**'

### The Europa Gateway

Now is a good time to think about how to make our data publicly available. Since none of the most common Web browsers supplies a Z39.50 client we must have a Web interface to query our installation with HTTP requests. Visit <http://europagate.dtv.dk/cgi-bin/egwcgi/80442/tform.egw> and complete the three first fields of the form. Leave the others to their default values. Press 'submit'. Now [search](#) for the nickname that you just gave your name server. Enjoy!

---

### Core skills

Anyone interested in setting up a vanilla-flavoured information gateway should be familiar with UNIX and its development environment in general. Knowledge of Perl-style regular expressions will make things a bit simpler. Programming skills and fluency in Perl are necessary for configuring an information gateway to fit a specific purpose, tuning performance and so on.

---

### Staff effort

Anyone who has the core skills listed above will be able to set up and configure a first gateway in under a week. With some experience it could be done in two hours. Experience shows that the maintenance of a gateway takes about four hours a week.

## References

---

### References

- Organizations
  - Danmarks Tekniske Videncenter & Bibliotek, DTV <http://www.dtv.dk>
  - Dublin Core Metadata Initiative <http://purl.oclc.org/dc/>
  - Free Software Foundation and the GNU Project <http://www.gnu.org>
  - Indexdata, Denmark <http://www.indexdata.dk/>
  - NetLab, Sweden <http://www.lub.lu.se/netlab/>
  - Sleepy-Cat Software <http://www.sleepycat.com/>
- Projects
  - All Engineering <http://www.lub.lu.se/eel/ae/index.html>
  - DESIRE <http://www.desire.org/>
  - EELS <http://www.lub.lu.se/eel/eelhome.html>
  - Europagate <http://europagate.dtv.dk>
- Software
  - BerkeleyDB <http://www.sleepycat.com/>
  - Combine <http://www.lub.lu.se/combine/>  
Mailing list [nwi@munin.lub.lu.se](mailto:nwi@munin.lub.lu.se)
  - Zebra <http://www.indexdata.dk/zebra/>
  - Z39.50 protocol <http://lcweb.loc.gov/z3950/>

## Credits

---

Chapter author: [Fredrick Rybarczyk](#)

With contributions from: Andy Powell, Jasper Tredgold

## 3.5. User profiles

### In this chapter...

---

- why profiles?
- personalisation
- characterising user interests
- authentication, trust and standards
- directory services
- legal issues

### Introduction

---

This chapter provides a brief overview of some issues surrounding the provision of personal profile services for Information Gateways. It is beyond the scope of this document to offer a comprehensive account of these complex issues. Instead, a brief summary of major points is provided alongside pointers to more detailed treatments available online.

### Why Profiles?

---

User profile services are a natural extension to the subject gateway approach. Subject-themed Information Gateways provide a focal point for broadly defined subject communities. Through the addition of user profile facilities, broadly-based gateways can begin to provide more specific 'views' into their information services. This is particularly important where a gateway's target audience includes multiple smaller communities. For example, a Social Science service such as [SOSIG](#) may have information appropriate for the Economics, Psychology and Law subject areas. Individuals in these professions may think of themselves as economists, psychologists or lawyers rather than as social scientists. A broadly based Social Science gateway that covers these topics (amongst others) might therefore benefit from an architecture which allows community specific or personalised views into a sub-set of the available resources. User profiles, which we might loosely define here as 'data structures that describe the properties of users', are an essential component of such a system since they allow a service to cross-reference information resources against user interests.

## Personalisation

---

The notion of a 'personalised' interfaces to Web content has become commonplace. There are challenges involved in the creation of such interfaces, but these typically share a common component: personal profiles. As used here, 'personal profiles' refers to the practice of describing individuals and various of their properties in a database for the purpose of improving their access to networked information resources. For example, a profile might store name and address details, home page URL, URL of an online image of that person, alongside details of their interests.

## Characterising User Interests

---

While there are no established standards for doing this, a simple guiding principle is to attempt to align the subject classification of documents and other 'discoverable' resources with user profile 'interest' classifications describing the subject or subjects that some user is interested in hearing about. For example, an information gateway targeted at the Economics community might adopt the JEL (Journal of Economic Literature) subject scheme both for user profiles and for classification. The SOSIG Grapevine service, similarly, has used the UDC subject scheme for personal interest profiling, to facilitate easy cross referencing with SOSIG catalogue records.

Many of the observations made in this handbook concerning the value of formal classification schemes and controlled vocabularies in the context of document description are also of relevance in the field of user profiling. There are, however, some differences. If complex structured vocabularies are to be used to allow users to describe their interests, a number of challenges arise for Information Gateway architects.

### **User Interface:**

There is a significant challenge associated with building an intuitive interface which allows users to pick subject headings from a (potentially very large) set of subject categories they interested in.

### **Multiple subject schemes:**

The problem of multiple classification schemes and mapping between them is as big a problem here as in document classification (See the section on controlled vocabularies).

### **Multiple interests:**

This is another potential usability problem. There is a case to be made for allowing users to define multiple 'profiles' for each of several potentially unrelated subject interests they may have. While this result can result in a 'cleaner' and more accurately structured profile, there is an associated cost in terms of the increased user interface complexity.

## Authentication, Trust and Standards

---

For an information gateway to offer personal profile based services, it is necessary for the service to have one or more mechanisms to establish the identity of users. There are a range of options here, from a simple stand-alone database of username/password pairs to more sophisticated cryptographic solutions. Gateway providers should be aware that there is as yet no widespread 'right answer' to this problem. Deployment of cryptographic (digital signature) technologies for this is at an early stage. Simpler username/password approaches (particularly when the default non-encrypted 'Simple authentication' HTTP authentication protocol is used) have their own problems. Users will frequently forget their passwords, and are known to be reluctant to go to the trouble of logging in to an authenticated service unless there is a clear benefit to doing so.

It is important to establish both formal and informal trust relationships with users when building a personalised, authentication-mediated Information Gateway. A formal 'privacy statement' for your service is a necessity. Users should know exactly what data you will be holding about them, and the purposes to which it will be put. The Platform for Privacy and Preferences (P3P) work of the World Wide Web Consortium (W3C) is a relevant standard here. P3P provides a common vocabulary for making such statements, both in simple natural language and in a machine-processable XML/RDF vocabulary. The idea here is to facilitate automatic negotiation between 'user agents' (i.e. Web browsers) and Web services such as information gateways.

The current Web model for acquiring user profile information from users usually involves the user completing a Web form. Users are often reluctant to do so, both due to lack of trust or knowledge regarding the remote service, or because it is simply a boring and repetitive task. The combination of metadata standards such as P3P, vCard and XML/RDF promises to make this task easier. vCard is a simple standard which specifies a common set of fields for personal profile data; in this sense it plays a similar role to that played by the Dublin Core element set in document description.

A P3P-aware browser and server should be capable of discussing, on behalf of their human counterparts (end user, service provider) the data fields requested by the server and the applications it will be acceptable to use these for. Whilst P3P is not yet widely deployed, Information Gateway services should be aware that such facilities are a likely development, and that their potential for service enhancements may be significant. For example, if P3P succeeds, Web services will be able to automatically ask for subject-interest information about users browsing their site.

#### EXAMPLE

##### P3P Example Scenario

The following example shows some of the current capabilities of the P3P data negotiation framework. P3P has an extensible architecture, and may in future versions allow such machine-processable statements to refer to arbitrary data structures (such as subject interest information).

**Note:** the english language text that follows has a precise mapping onto the formal, machine processable data structures defined in the P3P specification. The actual text below is based on an example from the AT&T Privacy Minder toolkit, and happens to fairly well characterise the current operating policies for the DESIRE project web site.

Sample (Fictional) P3P Privacy Statement for <http://www.desire.org/>

The DESIRE project makes the following statement for the Web pages at <http://www.desire.org/>

*We collect clickstream and user agent information stored in standard HTTP log files. We use this information for Web site and system administration. We do not distribute this information or use it in a way that would identify you.*

*We also have forms on our Web site where we may collect your contact information, information about your computer, demographic information, and information about your preferences. We use this information to complete transactions, provide customized services, and contact you. We may also use it for system administration and for research and development. We will not distribute this information to other organisations.*

*We use "cookies" on some of our Web pages in order to provide customize services to you and to research the way people use our web site.*

For further sample applications of P3P see AT&T's "Privacy Minder" tools at <http://www.research.att.com/projects/p3p/pm/>

## Directory Services

---

One possible technology applicable to Information Gateway user profile services is LDAP, or more broadly, 'white page' directory services. LDAP is a derivative of the older X.500 standard for representing personal data in a set of networked databases. LDAP does not address problems such as the classification of user interests, but does provide a widely implemented standard for representing name, address and contact detail information. Whether a directory-based approach, rather than a privately managed database, is appropriate will depend on the nature of your application. Where profile information will be exploited by a number of loosely connected Information Gateways, LDAP may be an attractive solution.

## Legal Issues

---

Any computer-based service which stores data about individuals should take legal advice about their practices, and in particular about the implicit or explicit contracts that they enter into with users. It is beyond the scope of this handbook to offer further guidance here, other than to say that the full complexities of the international environment of the Web have yet to be worked through in court. Different countries have varying laws regarding the management and storage of personal profile data; service providers should consequently proceed with caution when making such systems available to an international user base.

## Glossary

---

**LDAP** - Lightweight Directory Access Protocol  
**P3P** - Platform for Privacy and Preferences  
**RDF** - Resource Description Framework  
**XML** - Extensible Markup Language

## References

---

Grapevine, <http://www.grapevine.sosig.ac.uk/>

P3P, <http://www.w3.org/P3P/>

M. Wahl, T. Howes & S. Kille, *RFC 2251, Lightweight Directory Access Protocol (v3) (Internet Engineering Task Force, Network Working Group, December 1997)*.  
<ftp://ftp.isi.edu/in-notes/rfc2251.txt>

## Credits

---

Chapter author: [Dan Brickley](#)

## 3.6. Interoperability

### In this chapter...

---

- why interoperability is important for information gateways
- the role of protocols such as LDAP, Whois++ and Z39.50
- interoperability between metadata formats, metadata crosswalks and metadata registries
- content issues: cataloguing rules and classification schemes

### Introduction

---

No single information gateway will be able to describe each and every relevant Internet resource, even if it is limited to a relatively small subject area. Therefore, as the Internet continues to grow, gateways will need to co-operate (and interoperate) with each other to create distributed systems with wide geographical and linguistic coverage. Place (1999) suggests that the international library community is well placed to take up this challenge. She also notes that a collaborative network known as IMesh will provide an open forum for exchanging ideas and technology.

Indeed, the consistent use of existing standards and technologies already permits a large amount of inter-gateway collaboration. A lot of technical effort has gone into building interoperability between search protocols and metadata formats and into developing gateway software that is able to cross-search more than one gateway.

#### EXAMPLE

##### [IMesh](#)

IMesh provides an open forum and mailing list for exchanging ideas and technologies for promoting information gateways.

This chapter will not explain in technical detail how to implement interoperability features in a gateway, but will provide an overview of the various issues surrounding gateway interoperability.

## Background

---

In a computer science context the term 'interoperability' is used to refer to the transparent management of different applications and software. In an information gateway context, interoperability generally means one of two things:

- being able to search, browse and retrieve information from distributed gateways based on (broadly) the same technologies, protocols and metadata formats
- being able to search, browse and retrieve information from distributed gateways based on a variety of software solutions, search and retrieve protocols and metadata formats

These two different challenges require slightly different solutions. Where the same protocols and metadata formats are in use, ensuring interoperability is usually a matter of making sure that each gateway is set up in a consistent manner and has the correct interfaces. For example, it should be relatively easy to ensure that all services based on the Whois++ search and retrieve protocol (e.g. services based on the ROADS software toolkit) can be cross-searched. Interoperability, in these circumstances, becomes less of a technical problem and more a matter of the consistent use of metadata formats and their related content standards (e.g. cataloguing and subject indexing).

Where services are based on a variety of protocols and metadata formats, however, these non-technical problems remain - indeed, they are usually more difficult to solve - but additional technical layers will also need to be developed, involving the production of inter-protocol gateways, 'middleware' and metadata crosswalks.

In practice, however, information gateways tend to be based on a relatively small number of technologies, protocols and metadata formats, at least when compared with the whole information universe. This means that any work carried out on integrating several selected protocols and formats will be applicable in a number of different situations.

## Information gateways and interoperability

---

Ensuring that information gateways are interoperable will generally require the consistent application of available standards. There are four main 'standards-based' factors affecting interoperability among information gateways:

- the use of different search and retrieve (or indexing) protocols
- the use of different metadata formats
- differences in cataloguing standards
- differences in subject indexing schemes

### Protocols

Interoperability among information gateways requires the consistent use of relevant protocols. The most relevant protocols for gateways are LDAP, Whois++ and Z39.50.

### The Lightweight Directory Access Protocol (LDAP)

LDAP (cf. e.g. RFC 2251) was developed as a simple alternative to the ISO X.500 protocol, a directory access protocol designed for providing access to distributed information about people (names, email addresses, telephone numbers, etc). Accordingly, most existing applications of LDAP are so-called 'white pages' services. However, there is no reason why LDAP cannot be used for other services, including information gateways.

### EXAMPLE

#### [The Isaac Network](#)

The Isaac Network - an initiative of the Internet Scout Project based in the Computer Sciences Department at the University of Wisconsin-Madison - is using an LDAP directory for Dublin Core metadata records about resources (Roszkowski and Lukas, 1998; Lukas and Roszkowski, 1999).

### Whois++

The Whois++ protocol was originally developed for directory services, to operate as a simple (template-based), distributed and extensible information lookup service (RFC 1835). Its extensible



architecture, however, meant that its developers expected it to find applications in a number of other information service areas. Whois++ also provides a general architecture that is designed for the indexing of distributed databases and then applies that architecture to link together a multiple number of these Whois++ servers into a distributed, searchable wide-area directory service (RFC 1913). Unlike other directory protocols (e.g. X.500 or LDAP), Whois++ does not require a hierarchical representation of data space, but servers 'refer' the clients to other servers in a Whois++ 'mesh' (RFC 1914). Queries are routed through this mesh based on 'forward knowledge' held by one server about another. In Whois++, this forward knowledge is maintained using the Common Indexing Protocol (CIP).

CIP is a protocol used between servers in a network to facilitate query routing, the 'act of redirecting and replicating queries through a distributed database system towards the servers holding the actual results via reference to indexing information' (Allen and Mealling, 1997). It is not part of Whois++ and indeed can be used with other protocols such as LDAP. CIP is based upon the concept of index summaries or centroids. A centroid can be considered as a summary of the structured information in a given server; for example, it could be a simple inverted index of the information contained within a database's templates. This can then be used, for (e.g.) query routing within a distributed database.

#### E X A M P L E

##### **ROADS use of Whois++ and centroids**

The ROADS software (from version 1) uses the Whois++ protocol to query (and retrieve information from) distributed servers containing structured descriptions (ROADS templates) of Internet resources. In addition, ROADS (version 2) makes use of the centroid facility of Whois++ to facilitate query routing between servers. It may be worth while describing these technologies in more detail.

In a cross-searching context, a ROADS 'index server' will periodically visit ROADS-based information gateways and generate an index summary (or centroid). The centroid for each service (or server) will contain all relevant index terms in that database, so that an initial search of the index server will determine which of the subject services has information that matches a given query. If desired, the query can then automatically be passed on to all the information gateways whose centroids indicate the existence of relevant index terms and the templates containing them returned for display to the end-user. Demonstrations of ROADS cross-searching are currently available on the Web (ROADS project, 1998), as are more detailed descriptions of the technologies that underlie it (e.g. Knight and Hamilton, 1995; Kirriemuir, et al., 1998).

- [ROADS](#)

#### **Z39.50**

The Z39.50 protocol (e.g. Library of Congress, 1999) is a standard for information retrieval approved by the National Information Standards Organization (NISO) - a committee accredited by the American National Standards Institute (ANSI). It has also been recognised by the International Organization for Standardization (ISO), where it is known as ISO 23950:1998.

The Z39.50 protocol allows client applications to search databases on remote 'target' servers and to retrieve relevant information. It therefore supports the retrieval of information from distributed remote databases (Turner, 1995). The first applications using it, for example software for distributed searching of library online public-access catalogues, were developed specifically for bibliographic data, but attribute sets can be defined to allow the protocol to work with many other types of data. For example, systems using Z39.50 have been developed for libraries, archives, museums and data archives.

#### E X A M P L E

##### **The AHDS gateway**

The Arts and Humanities Data Service (AHDS) consists of five distributed subject-based service providers which, in addition to their other responsibilities, provide access to descriptions of digital resources in five separate subject domains:

- Archaeology Data Service (ADS)
- History Data Service (HDS)
- Oxford Text Archive (OTA)

- Performing Arts Data Service (PADS)
- Visual Arts Data Service (VADS)

Each of these services operates within a resource description context specific to its own subject domain. For example, the Oxford Text Archive - a service provider for literary and linguistic texts - would normally describe resources using a metadata format known as 'Text Encoding Initiative (TEI) headers'.

The AHDS has implemented a resource discovery system which provides unified access to these heterogeneous (and distributed) resource descriptions using Dublin Core and a Z39.50 gateway (Miller and Greenstein, 1997). Greenstein and Murray (1997, p. 56) explain:

[The Z39.50-based] software acts as a mediating layer between on the one hand, a World Wide Web interface from which users query a range of different catalogue databases and to which merged result sets are returned to the user, and on the other, the underlying catalogue databases themselves. From the users point of view, this 'middleware' irons out any differences that may exist in the underlying databases (e.g. in their native record structure, query language, and record syntax).

- [AHDS gateway](#)

Z39.50 has not been widely implemented by information gateways. However, there is a wider need to ensure that gateways can interoperate with other resource discovery systems (such as library OPACs, hybrid library systems) and different metadata formats. For these reasons, projects like ROADS have needed to address issues relating to gateway interoperability with Z39.50.

#### E X A M P L E

##### **ROADS (Whois++) interaction with Z39.50**

Although ROADS databases normally make resource descriptions available using Whois++, the ROADS project realised that in some situations it would be desirable to make such databases available to end-user client and intermediate systems that use the Z39.50 protocol.

Two main approaches were adopted:

1. A Z39.50 to Whois++ gateway. In this solution, the gateway functions as a Z39.50 server, accepting queries from Z39.50 client systems. It then converts them to Whois++ queries and passes them to the ROADS server. As the ROADS server returns results, they are converted into a suitable format for use by Z39.50 client systems and returned to the client as a Z39.50 results set. A Z39.50 to Whois++ gateway, known as ZEXI, has been developed as part of the ROADS project. It is based on the Isite Information System available from CNIDR. ZEXI returns simple, unstructured text-based records known as SUTRS.
2. Loading ROADS records into a Z39.50-based database. The second approach involves copying records from a ROADS database into another database that has a Z39.50 interface. Typically, the records will require some form of conversion during the copying procedure. Candidate Z39.50 database systems include Isite and the Zebra System developed by Index Data. The Zebra Z39.50 server can make converted ROADS records available in two structured formats (USMARC and GRS-1) and in an unstructured format (SUTRS).

Documentation (and software) on making ROADS databases accessible using this second approach (the ROADS Z39.50 Plugin) is available from the ROADS project Web pages.

- [ROADS Z39.50 plugin](#)

#### **Metadata formats**

##### **Metadata crosswalks**

Different information gateways will often use different metadata formats. For this reason there is a need for crosswalks (or mappings) between formats that can be used as the basis of interoperable

systems (such as middleware) or for conversion programs.

## CROSS REFERENCE

### [Metadata formats](#)

A number of inter-metadata crosswalks exist, many based on Dublin Core (RFC 2413). Core metadata formats are well placed to act as intermediaries for semantic interoperability between heterogeneous resource description models. Weibel (1997, p. 18) suggests that the promotion of a 'commonly understood set of core descriptors will improve the prospects for cross-disciplinary search by unifying related attributes'. He additionally suggests that an important approach to interoperability in a heterogeneous resource description environment would be to map many description schemas into a common set (such as Dublin Core) which would give users 'a single semantic model for searching'.

A number of Dublin Core (DC) based mappings currently exist; for example, there are important crosswalks from Dublin Core to USMARC (Caplan and Guenther, 1996; Network Development and MARC Standards Office, 1997). Other people and organisations have also produced DC mappings for various other formats including TEI headers, the Nordic MARC formats (as part of the Nordic Metadata Project) and UNIMARC (for project BIBLINK). A collection of these metadata mappings is maintained by Day (1996).

The ROADS project has produced metadata crosswalks between ROADS templates, Dublin Core, SOIF and the USMARC format.

### **Metadata Registries**

Metadata formats require consistent application. This is particularly a problem with formats that are easily adaptable and extensible, such as ROADS templates or Dublin Core. It would be possible for an information gateway to modify (or customise) a metadata format so much that the service based on it would no longer be interoperable (cross-searchable) with other gateways.

One solution would be to require all gateways to conform to an agreed set of metadata attributes. However this goes against the very flexibility that gateways require in order to provide a good service to their own users. What is needed is a way of recording current practice so that gateways can modify metadata formats in the knowledge of what other gateways have done and without the problem of 'reinventing the wheel'.

## E X A M P L E

### **The ROADS Template Registry**

ROADS templates are defined for 15 different resource types. These are known as template types. Some of these template types (e.g. DOCUMENT, MAILARCHIVE and SERVICE) originate in the original IAFA template specification (Deutsch et al. 1994). Other templates have been developed specifically for ROADS-based services (e.g. PROJECT). At least one of the others (TRAINMAT, for training materials) was independently developed and has been published as RFC 2007.

Each template type has a number of set attributes. Some of these are specific to one template type, others are not. ROADS templates use what the IAFA specification calls 'clusters' to group together information on names, addresses and other contact details. Clusters currently in use describe a USER (an individual) or an ORGANIZATION. ROADS-based services can also add new attributes and create new template types.

Experience with ROADS-based gateways demonstrated a need for a metadata registry. The creation of new template types and the adaptation and extension of existing template types by subject services meant that there was no central location where the latest forms of these could be recorded.

The ROADS Template Registry takes the form of a list of template types, including all metadata attributes that have been proven to be useful. The aim of the registry is to preserve flexibility - to allow the creation of new template types and attributes where necessary - but also to prevent the unnecessary proliferation of template types and attributes and to maintain some level of consistency.

Consistency is extremely important in the context of ROADS cross-searching and interoperability. It would be possible for a ROADS user to consider creating a new template type

for (say) recorded music; it would be desirable to base this on an existing template type (e.g. VIDEO) and to use - wherever possible - attributes and clusters that are common to more than one existing template type.

- [ROADS template registry](#)

What are needed are extensible metadata registries which provide canonical definitions of all elements and also disclose local uses. These registries should be understandable by both humans and machines. ISO/IEC 11179:1997 - Specification and standardization of data elements is a formal standard for expressing the semantics of data elements suitable for registries, but few metadata registries based on this standard currently exist.

#### EXAMPLE

##### ISO/IEC 11179 registries

###### Environmental Data Registry (EDR)

The U.S. Environmental Protection Agency (EPA) developed its Environmental Data Registry (EDR) as a comprehensive and authoritative source of reference information about environmental data. The registry provides information on data names, definitions, formats, and relationships and identifies organisations (or individuals) responsible for the various data. Registered users can also register new data elements in the EDR.

- [EDR](#)

###### National Health Information Knowledgebase (NHIK)

The Australian Institute of Health and Welfare (AIHW) developed its National Health Information Knowledgebase (NHIK) as an 'electronic repository' for health metadata. Data elements within the Knowledgebase have been documented using ISO/IEC 11179.

- [NHIK](#)

#### Content issues

##### Cataloguing

In practice, interoperability is not just dependent upon consistency in the use of the metadata format itself but is also dependent upon the consistency of the content contained within the format. For example, in the library community the MARC formats specify a framework for the description of bibliographic items while the content of MARC records will often conform to other standards, usually based on one of the International Standard Bibliographic Descriptions (ISBDs) or cataloguing rules derived from them.

For this reason, the formulation of cataloguing guidelines will be an important part of the interoperability strategy of a gateway (e.g. Day, 1998). This will mean taking account of cataloguing practice in other gateways and the production of standardised cataloguing rules, considering such issues as:

- chief sources of information
- capitalisation
- date formats
- language codes
- formats for personal and corporate names

#### CROSS REFERENCE

[Cataloguing](#)

#### Subject classifications

Another content-based area where interoperability is likely to become an issue is in the application of subject information in the form of classification schemes and thesaurus terms.

Classification schemes provide an information gateway with a browsing structure. It is possible that two or more distributed gateways could be combined to form a single service. Successful cross-browsing will depend upon the consistent application of the same classification scheme. Therefore, information gateways that want to facilitate cross-browsing should, wherever possible, use the same classification system.

Otherwise, complex mappings will have to be produced to enable conversion between schemes. This may not be too difficult at the higher levels of a universal subject hierarchy but where any detail is involved it will become problematic because of theoretical, conceptual, cultural and practical differences between systems.

### CROSS REFERENCE

[Subject indexing and classification](#), [Co-operation between gateways](#)

## Conclusions

---

It is important for all information gateways to consider interoperability issues. It is generally agreed that the way forward for information gateways is increased co-operation; successful information gateway co-operation will depend upon successful interoperability and in the consistent application of standards regarding such matters as protocols, metadata formats, cataloguing rules and subject classification schemes. Gateways can start to make immediate use of existing tools that promote interoperability and to build the technical links between distributed gateways that will form the basis of any future international co-operation.

## Glossary

---

**ADS** - Archaeology Data Service  
**AHDS** - Arts and Humanities Data Service  
**AIHW** - Australian Institute of Health and Welfare  
**ANSI** - American National Standards Institute  
**CIP** - Common Indexing Protocol  
**CNDR** - Center for Networked Information Discovery and Retrieval  
**EDR** - Environmental Data Registry  
**EPA** - Environmental Protection Agency  
**HDS** - History Data Service  
**IAFA** - Internet Anonymous FTP Archive  
**IEC** - International Electrotechnical Commission  
**IETF** - Internet Engineering Task Force  
**ISBD** - International Standard Bibliographic Description  
**ISO** - International Standards Organization  
**LDAP** - Lightweight Directory Access Protocol  
**MARC** - Machine-Readable Cataloguing  
**NHIK** - National Health Information Knowledgebase  
**NISO** - National Information Standards Organisation  
**OTA** - Oxford Text Archive  
**PADS** - Performing Arts Data Service  
**RFC** - IETF Request for Comments  
**ROADS** - Resource Organisation and Discovery in Subject-based services  
**SUTRS** - Simple Unstructured Text Record  
**TEI** - Text Encoding Initiative  
**UNIMARC** - Universal MARC format  
**VADS** - Visual Arts Data Service  
**Whois++** - A 'lightweight' Internet protocol for information retrieval  
**X.500** - An ISO directory protocol  
**Z39.50** - An ANSI/NISO developed protocol for information retrieval - also known as ISO 23950

## References

---

- AHDS gateway, [http://ahds.ac.uk:8080/ahds\\_live/](http://ahds.ac.uk:8080/ahds_live/)
- EDR, <http://www.epa.gov/edr/>
- IMesh, <http://www.desire.org/html/subjectgateways/community/imesh>
- Isaac Network, <http://scout.cs.wisc.edu/research/index.html>
- NHIK, <http://www.aihw.gov.au/services/health/nhik.html>
- ROADS, <http://www.ilt.bris.ac.uk/roads/>
- ROADS template registry, <http://www.ukoln.ac.uk/roads/templates/>
- ROADS Z39.50 plugin, <http://www.ilt.bris.ac.uk/roads/software/zplugin/>
- J. Allen & M. Mealling, *The architecture of the Common Indexing Protocol (CIP) (FIND Working Group, Internet-Draft, 18 November 1998)*.  
<ftp://ftp.isi.edu/internet-drafts/draft-ietf-find-cip-arch-02.txt>
- P. L. Caplan, & R. S. Guenther, 'Metadata for Internet resources: the Dublin Core Metadata Element Set and its mapping to USMARC', *Cataloging and Classification Quarterly* 22 nos. 3-4 (1996), 43-58.
- M. Day, *Mapping between metadata formats (Bath: UKOLN The UK Office for Library and Information Networking, 1996)*.  
<http://www.ukoln.ac.uk/metadata/interoperability/>
- M. Day, *ROADS cataloguing guidelines (Bath: UKOLN The UK Office for Library and Information Networking, 1998)*.  
<http://www.ukoln.ac.uk/metadata/roads/cataloguing/cataloguing-rules.html>
- P. Deutsch, A. Emtage, M. Koster & M. Stumpf, *Publishing information on the Internet with Anonymous FTP (Internet Engineering Task Force, Internet Draft, September 1994)*.  
<http://info.webcrawler.com/mak/projects/iafa/iafa.txt>
- P. Deutsch, R. Schoultz, P. Faltstrom & C. Weider, *RFC 1835, Architecture of the WHOIS++ service (Internet Engineering Task Force, Network Working Group, August 1995)*.  
<ftp://ftp.isi.edu/in-notes/rfc1835.txt>
- P. Faltstrom, R. Schoultz & C. Weider, *RFC 1914, How to interact with a Whois++ Mesh (Internet Engineering Task Force, Network Working Group, February 1996)*.  
<ftp://ftp.isi.edu/in-notes/rfc1914.txt>
- J. Foster, M. Issacs & M. Prior, *RFC 2007, Catalogue of network training materials (Internet Engineering Task Force, Network Working Group, October 1996)*.  
<ftp://ftp.isi.edu/in-notes/rfc2007.txt>
- D. Greenstein & R. Murray, 'Metadata and middleware: a systems architecture for cross-domain discovery' in P. Miller & D. Greenstein, eds., *Discovering online resources across the humanities: a practical implementation of the Dublin Core (Bath: UKOLN on behalf of the Arts and Humanities Data Service, October 1997)*, 56-62.  
[http://ahds.ac.uk/public/metadata/disc\\_06.html](http://ahds.ac.uk/public/metadata/disc_06.html)
- ISO 23950:1998, *Information and documentation - Information retrieval (Z39.50) - Application service definition and protocol specification (Geneva: International Organisation for Standardization, 1998)*.
- ISO/IEC 11179:1997, *Information technology - Specification and standardization of data elements (Geneva: International Organisation for Standardization, 1997)*.
- J. Kirriemuir, D. Brickley, S. Welsh, J. Knight & M. Hamilton, 'Cross-searching subject gateways: the query routing and forward knowledge approach', *D-Lib Magazine* (January 1998).

<http://www.dlib.org/dlib/january98/01kirriemuir.html>

J. P. Knight & M. Hamilton, *Overview of the ROADS software (LUT CS-TR 1010. Loughborough: Loughborough University of Technology, Department of Computer Studies, 1995).*

<http://www.roads.lut.ac.uk/Reports/arch/arch.html>

Library of Congress, *Z39.50 Maintenance Agency [home page], (Washington, D.C.: Library of Congress 1999).*

\*\*\* URL needed?

C. Lukas & M. Roszkowski, *'The Isaac Network: LDAP and distributed metadata for resource discovery', Third IEEE Meta-data Conference, National Institutes of Health, Bethesda, Md., USA, 6-7 April 1999.*

<http://computer.org/conferen/proceed/meta/1999/papers/46/clukas.html>

P. Miller & D. Greenstein, *Discovering online resources across the humanities: a practical implementation of the Dublin Core (Bath: UKOLN on behalf of the Arts and Humanities Data Service, October 1997).*

<http://ahds.ac.uk/public/metadata/discovery.html>

Network Development and MARC Standards Office, *Dublin Core/MARC/GILS Crosswalk (Washington, D.C.: Library of Congress, 4 July 1997).*

<http://lcweb.loc.gov/marc/dccross.html>

E. Place, *'International collaboration on Internet subject gateways', 65th IFLA Council and General Conference, Bangkok, Thailand, 20-28 August 1999.*

<http://www.ifla.org/IV/ifla65/papers/009-143e.htm>

ROADS project, *CrossROADS (Bath: UKOLN The UK Office for Library and Information Networking, 1998).*

<http://roads.ukoln.ac.uk/crossroads/>

M. Roszkowski & C. Lukas, *'A distributed architecture for resource discovery using metadata', D-Lib Magazine (June 1998).*

<http://www.dlib.org/dlib/june98/scout/06roszkowski.html>

F. Turner, *An overview of the Z39.50 Information Retrieval standard (UDT Occasional Paper, 3. Ottawa: IFLA Universal Dataflow and Telecommunications Core Programme, 1995).*

<http://www.ifla.org/VI/5/op/udtop3.htm>

M. Wahl, T. Howes & S. Kille, *RFC 2251, Lightweight Directory Access Protocol (v3) (Internet Engineering Task Force, Network Working Group, December 1997).*

<ftp://ftp.isi.edu/in-notes/rfc2251.txt>

S. Weibel, J. Kunze, C. Lagoze & M. Wolf, *RFC 2413, Dublin Core metadata for resource discovery (Internet Engineering Task Force, Network Working Group, September 1998).*

<ftp://ftp.isi.edu/in-notes/rfc2413.txt>

C. Weider, J. Fullton & S. Spero, *RFC 1913, Architecture of the Whois++ Index Service (Internet Engineering Task Force, Network Working Group, February 1996).*

<ftp://ftp.isi.edu/in-notes/rfc1913.txt>

## Credits

---

Chapter author: [Michael Day](#)

With contributions from: Rachel Heery

## 3.7. Scalability

### In this chapter...

---

- an overview of scalability issues
- user interface and usability
- administration and management
- systems issues

### Introduction

---

Scalability is an issue that needs to be considered when designing any system for long-term data storage. It is not sufficient to design your system to meet current requirements; you also need to take into account (or at least be aware) how your collection of data is likely to grow in the coming years. A system that is perfectly adequate for storing, manipulating and providing access to a small number of records or data may be quite unable to cope if the amount of data increases by one or two orders of magnitude.

This chapter will look at the problems and issues specific to subject gateways that arise because of such increases in database size and will consider approaches to dealing with these problems.

### Background

---

At present, subject gateways tend to consist of no more than a few thousand records because of the manual effort required to select and catalogue Internet resources. Even a 'large' subject gateway typically has only about six or seven thousand records. This is very small in comparison with traditional online bibliographic databases. Consequently, the problems associated with storing and retrieving large collections of bibliographic data, such as recall and precision in searches and search engine functionality, have not yet been significant.

It seems unlikely that individual subject gateways are capable of growing significantly in size, given current funding models. Only directories that have limited or no quality criteria, high levels of funding or possibly voluntary effort - such as Yahoo!, OCLC's NetFirst or the Open Directory Project - seem to be capable of producing manually-created databases with sizes of the order of hundreds of thousands of records.

The likely method of growth for subject gateways seems instead to be via collaborative effort. There are two approaches to building a collaborative subject gateway. The first is for a number of different organisations to contribute records to a central database. The problems with such an approach are likely to be concerned with the size of the database, maintaining reasonable performance on a single machine and providing network access to it. The second approach is for each organisation to maintain its own database, allowing the end-user to search across one or more of them depending on the nature of their query. In some cases a combination of the two approaches may be appropriate. These methods allow a real or virtual increase in size of the collection of resources presented to the end-user.

#### CROSS REFERENCE

[Interoperability, Co-operation between gateways](#)

We have also begun to see the creation of harvesting software which enables the automated indexing of Internet resources whilst retaining a degree of quality because of the ability to choose the seeding URIs for the robot. The first phase of the DESIRE project developed some harvesting tools that can be used in conjunction with the ROADS and Zebra software. Such mechanisms have the potential to create databases at least one order of magnitude larger than those of current gateways. This increase in size of the database presented to the end-user and the ability to pass a single search to a number of different databases produce new problems that need to be addressed.



## E X A M P L E

### Case study - SOSIG Link Harvester Index

The SOSIG Link Harvester Index is an online database separate from the main SOSIG Internet Catalogue. Whereas the resources found in the SOSIG Internet Catalogue have been selected manually by subject experts, those in the SOSIG Link Harvester Index have been collected by software called a harvester (similar mechanisms may be referred to as robots or Web crawlers). The records in the Internet Catalogue provide the list of seeding URLs for the harvester.

- [Combine](#)

### CROSS REFERENCE

[Harvesting, indexing and automated metadata collection](#)

Experiments are also taking place using useful 'lists of lists', not normally added to the catalogue, as seeding URLs.

**Note:** problems with large subject gateway databases are not limited to the user interface - the SOSIG Link Harvester Index has already had to be limited to 50,000 records because of indexing limitations in the ROADS software.

## Scalability Issues

---

### Overview

Part of the scalability problem is concerned with interface and usability issues. These include the presentation of large results sets to the user, the means by which the cross-search paradigm is presented and the ranking or filtering of any results produced. Another part of the problem is concerned with the management of such collections: for example, the need for automated mechanisms for link checking and perhaps for detecting changes to sites that require their descriptions to be updated. Finally there are issues relating to the computer systems used to run the subject gateway service, such as the need for databases that can handle much larger collections of data.

The rest of this chapter therefore consists of three sections; the first will look at user interface and usability issues, the second will consider administration and management issues and the third will consider the systems issues involved in maintaining large collections of records.

### User interface and usability issues

With a relatively small database, the issue of precision in searching is not very important, since the user can scroll quickly through a results set to discover which are the most useful records. However, as the size of the database increases, so does the average number of records retrieved, and it then becomes much more difficult to select the most relevant and useful ones. This problem can be approached in two ways:

- by increasing the precision of the search so that fewer irrelevant results are returned
  - by ranking and filtering the results set so that the most relevant results stand out in some manner
- Mechanisms for increasing precision of searches

Here are some ways in which the precision of searches can be increased:

1. Allow searching by individual fields, such as title, as a way of increasing the usefulness of the search terms. Fields containing 'extra' information such as geographical area or type of resource will also be helpful for sorting relevant from irrelevant information.
2. Allow the use of keywords. Keywords may be added to records as a means of describing the main topics dealt with in the resource being catalogued. This generally increases the 'recall' of searches. However, if keywords are combined with fielded searching, so that the keyword field can be specified, the precision of the results can also be improved.
3. Allow the use of controlled vocabularies. These serve mainly to improve the recall of keyword searches and are usually organised into hierarchical structures, making it easier for the user to find the most relevant and specific term. Keyword searching using controlled vocabularies may cause problems with cross-searching, however, and requires the cross-searched catalogues to use the same vocabularies or to have a cross-mapping scheme

drawn up for them.

#### CROSS REFERENCE

[Subject indexing and classification](#)

### Displaying large results sets

Typically, large results sets cannot be displayed on a single Web page. This is because of the time taken to retrieve the data and because of scrolling problems for the end-user. The ROADS software limits the total number of records which can be returned by a search but, as the size of the database increases, the proportion of searches resulting in 'too many hits' will also increase. In addition to reducing the number of hits returned, by increasing the precision of searches, it may also be sensible to investigate mechanisms for improving the way in which records are displayed. These may include:

1. Limiting the number of records displayed at a time (note that ROADS doesn't currently support this feature). Remember that end-users may still not look through many pages of results even when they are presented in small chunks.
2. Ranking and/or filtering the results. It may be possible to use metadata both to rank and filter results, for example to display results only for resources that are of undergraduate level or above. Such a technique could also be combined with recommendations (quality ratings) from other people in the end-user's subject area. A detailed discussion of these techniques is beyond the scope of this chapter; however some work in this area is currently under way in the DESIRE II project.

#### CROSS REFERENCE

[Quality selection: Quality ratings](#)

### Browsing larger collections (including cross-browsing)

Most subject gateways provide a browsing interface to their data in addition to a search interface. Many of the issues raised above apply equally to the browse interface. For example, as the number of records in the database grows, the lists of records presented in the browse interface are likely to become too long to be shown on a single Web page.

The browse interface is typically designed (at least in part) around the controlled vocabulary (classification scheme) for keywords described above. As the database increases in size, the number of records per section will also increase unless the granularity of the classification scheme is increased. Therefore, there are some design decisions that need to be taken concerning the depth and complexity of the classification scheme used.

#### CROSS REFERENCE

[Subject indexing and classification, User interface implementation](#)

It is worth noting that a combination of browse and search interfaces may help the end-user. This may be achieved by embedding a restricted search interface into each sub-section of the browse interface, returning results that are only applicable to that sub-section.

### Administration and Management Issues

As the number of records in a subject gateway database increases, the techniques used to manage it may need to change. Manual checking of records is likely to be feasible for a small database, but who wants to check 7,000 records by hand? What about 50,000 records?!

Some areas where automated checking of records may be possible are:

1. Link checking. The ROADS software provides an automated link checker which will confirm the validity of the URLs in all the records in a subject gateway's database on a regular basis.
2. Resource updates. There is a danger that the descriptions of resources held in subject gateways will become out of date as the resources themselves are updated. It may be possible to develop robot-based tools that check for potentially 'significant' changes to the resources described in a subject gateway's database, automatically warning resource cataloguers of the records that are likely to need updating.
3. Review-by dates. By embedding a 'review-by' date into every resource description you can

be notified automatically that a record hasn't been checked recently. Note that ROADS supports this feature out of the box.

## CROSS REFERENCE

[Collection management](#)

### Systems Issues

It is clear that as a database grows the amount of disk space it requires will also grow. Memory and CPU power requirements will probably also increase. It is possible that database software that copes with 10,000 records may not cope efficiently with 100,000 records. For example, there is some evidence that the file system based database software supplied with ROADS by default does not cope well with databases larger than about 50,000 records. In theory, ROADS allows you to plug in alternative back-end databases. However, it is not clear how many services are actively using this feature.

There may also be performance problems associated with cross-searching large numbers of large databases. The searching system has to wait for results to come back from all the databases that it is searching. This may tie up network and other resources on that system. Research is currently being done within the DESIRE project into the areas of parallel searching and results interfaces which return results to the user as and when they become available. Findings in this area will be published on the [DESIRE Web site](#).

### Glossary

---

**DESIRE** - Project funded under the European Union's Telematics for research Programme to enhance and facilitate Web usage among researchers in Europe (producer of this handbook)

**OCLC** - Online Computer Library Centre Inc.

**ROADS** - ROADS is a set of software tools to enable the set up and maintenance of Web based subject gateways.

**SOSIG** - The Social Science Information Gateway

### References

---

Combine, <http://www.lub.lu.se/combine>

DESIRE, <http://www.desire.org/results/training/D8-2af.html>

OCLC, <http://www.oclc.org/>

Open Directory Project, <http://dmoz.org/>

SOSIG Harvester, <http://www.sosig.ac.uk/roads/cgi/search.pl?form=harvester>

Yahoo!, <http://www.yahoo.com/>

### Credits

---

Chapter author: [Phil Cross](#), [Andy Powell](#)

## 3.8. Future proofing

### In this chapter...

---

- importance of planning for the future
- planning ahead: hardware, software and content

## Introduction

---

It is in the interests of all associated with the service to make reasonable attempts to future proof investment in the subject gateway. In this chapter we will consider how concern for future proofing can influence the gateway's decisions as regards hardware, software and content. Good decisions in these areas will provide a sound foundation for the future of the gateway. We will give a brief overview of some issues related to planning for the future in an area of rapid technological change and introduce some thoughts on how planning relates to decision making in the context of subject gateways.

The continued existence of a gateway depends ultimately on a sound business model with assured income. The wider aspects of business planning and marketing will be dealt with elsewhere. Issues relating to system requirements and scalability are also dealt with in more detail in other chapters. In this chapter we will relate planning and decision making to the specific areas of software, hardware and content.

### CROSS REFERENCE

[System requirements overview](#), [Scalability](#)

## Background

---

Different gateways will have different strategic objectives which will be expressed in the key characteristics of the services they provide and the level of innovation to which they aspire. Some gateways may wish to deliver services using the latest technology and to gain a reputation for introducing new features and incorporating the most recent software developments; other gateways may be more concerned with inter-working with legacy technology and content and may regard leading-edge technology as inappropriate. Some gateways will want to spend resources on research and development work, while others may want to identify reliable existing products.

Whatever the objectives of the gateway, some general principles can be identified which should inform decision making.

## Key factors for decision making

---

The gateway's decisions regarding hardware, software or content must take into account various imperatives. Each gateway must identify its own specific criteria and these criteria will differ depending on the gateway's priorities. However, there are some generic principles underlying the process of decision making which may be considered to be common to all gateways:

### **1. Planning for change.**

Search services are a growth area in the fluid Internet environment. This area is characterised by rapid shifts; new products are coming onto the market, new gateways are being set up and new technologies and standards are being developed. In addition the sectors in which gateways are working (education, libraries, knowledge industries) are also subject to change. Gateways need to be aware of new opportunities offered by change and be flexible enough to exploit them. In practical terms, this may mean delivering services to new audiences, incorporating new data structures, inter-working with services which may be based on different technologies. It may mean migrating to new systems, merging with other services, or taking on new service areas.

### **2. Decisions need to be based on criteria that are aligned with the gateway's strategic objectives.**

The gateway's strategic objectives need to be realised in day-to-day decisions. This means that all staff in the gateway need to be aware of the objectives and how they relate to their own decisions. For example, the choice of hardware needs to be informed by plans for growth, the choice of software must take account of the costs of inter-working with other services and the choice of metadata standards depends on users' search requirements and on the cost limitations for metadata creation.

### CROSS REFERENCE

[System requirements overview](#), [Metadata formats](#)

### **3. Taking account of the environment.**

Decisions need to be informed by knowledge of the environment. Who are the ultimate users of the service and what are their requirements now? How will their needs change? What are the priorities of the investors (funding bodies) and how can they be influenced? Who are the competitors? What are the differentials that distinguish your gateway?

The gateway will need to be aware of the effect of changes in the environment so that it can position itself to take advantage of opportunities, for example in the following ways:

- the system needs to adapt to new methods of data creation, new methods of service delivery
- modular design so that the system can change incrementally

## Conclusion

---

Sound decisions regarding system and content will contribute to future proofing the gateway. However, lasting success depends on many factors outside the control of the gateway itself. Future proofing needs to be seen as just one part of the wider strategic planning process which gateways need to undertake.

## Credits

---

Chapter author: [Rachel Heery](#)

---

Return to:  
[Handbook Home](#)  
[DESIRE Home](#)

[Search](#) | [Full Glossary](#) | [All References](#)

Last updated : 26 April 00

[Contact Us](#)  
© 1999-2000  
[DESIRE](#)