Gateway Standardization: A Quality Assurance Framework for Metadata

Brian Kelly, Amanda Closier, and Debra Hiom

Abstract

As digital library services develop from project demonstrators to mature, mission-critical services, it becomes necessary to develop and implement systematic procedures that will ensure the quality of the content, the functionality of the service, accessibility to a wide range of users and devices, and interoperability with other services. This article describes a quality assurance methodology that has been developed to support digital library programs in the United Kingdom higher and further education sectors. The article describes the approaches taken by the SOSIG subject gateway service in developing and maintaining a national service that is dependent on quality metadata. The article then outlines a quality assurance framework, which has been developed to support the Joint Information Systems Committee's (JISC) digital library programs in the UK and its application to metadata. The article concludes by describing a self-assessment toolkit that can be used by service providers to ensure that they have addressed the key areas.

The Web has now established its importance for providing access to scholarly resources in teaching and research. As digital library services develop from project demonstrators to mature, mission-critical services, it becomes necessary to develop and implement systematic procedures that will ensure the quality of the content, the functionality of the service, accessibility to a wide range of users and devices, and interoperability with other services. In the UK we have been working toward this end by developing a "quality

assurance (QA) methodology" to support digital library programs in the UK higher and further education sectors. This article describes the approaches taken toward developing and maintaining a national service that is dependent on quality metadata. The self-assessment toolkit we have developed can be adopted by subject gateway service providers to ensure that they have addressed the important issues facing digital library services—standardization and quality control.

Background

In the UK the Joint Information Systems Committee (JISC), which funds a range of networked services for the higher and further education communities, has played a key role in the development of digital library services. The JISC established the eLib program (eLib, 2001) in the mid-1990s, providing an opportunity for experimentation in multiple areas, including a strand for the establishment of pilot subject gateways. Following the success of the eLib program and the recognition of the Web as the key delivery platform for scholarly resources, the JISC subsequently established a strategy for accessing these resources seamlessly. Initially known as the DNER (Distributed National Electronic Resource) but later renamed the JISC Information Environment (IE), the implementation of this strategy is based on a number of JISC programs that fund the development of a wide range of projects. These projects will, together with related JISC service developments, help to provide the IE's content and technical infrastructure.

An example of one of JISC's national services is the Resource Discovery Network (RDN), which provides access to scholarly resources in various subjects. The RDN is an ambitious subject gateway system made up of eight area (or hub) subject gateways. These services (as indicated below) are hosted at particular universities throughout the UK and draw upon the expertise of over seventy educational and research organizations, including the Natural History Museum and the British Library. A summary of the RDN hubs is given in Table 1.

The RDN is now recognized as one of the Web's most reputable scholarly resources, with clear missions and interfaces set in place. Although, inevitably, there will continue to be a need for experimentation as new formats and protocols are developed and different types of services are evaluated, there is now a need to ensure that project deliverables can be deployed into a service environment with ease. In other words, once a gateway is built, it is necessary to establish systematic maintenance procedures, as well as continue to add resources to it. What follows is a description of a quality assurance (QA) framework for maximizing digital library services.

Table 1. Examples of Resource Discovery Network Hubs

Service	Area	Host
ALTIS	Hospitality, Leisure, Sport, and Tourism	University of Birmingham
Artifact	Arts and Creative Industries	Manchester Metropolitan University
BIOME	Health and Life Sciences	University of Nottingham
EEVL	Engineering, Mathematics, and Computing	Heriot Watt University, Edinburgh
GEsource	Geography and Environment	Consortium of Academic Libraries in Manchester
HUMBUL	Humanities	University of Oxford
PSIgate	Physical Sciences	Consortium of Academic Libraries in Manchester
SOSIG	Social Sciences, Business, and Law	University of Bristol

Quality Assurance Framework

The SOSIG case study (explained in detail below) outlines a practical approach for ensuring the quality of the service's metadata and hence maintaining the quality of the service. With the success of SOSIG's quality assurance procedures, it became clear that these methods could be implemented on a wider scale—to other JISC-funded services. In this section we describe how the JISC has funded the development of a quality assurance methodology for its digital library programs and how this methodology can be applied to the creation and management of metadata.

QA Focus

In 2001 the JISC issued a call for a "Digitisation and QA Focus" service (Joint Information Systems Committee, 2001). The call recognized that "Past digitisation programmes tended to operate in an environment where technologies were relatively immature and unstable, therefore suggesting a research-orientated approach to the management of digitisation activity." Following a successful bid the project (which was renamed "QA Focus") was provided initially by UKOLN (a national center of expertise in digital information management based at the University of Bath) in conjunction with the Institute of Learning and Research Technology (ILRT) based at the University of Bristol (ILRT is the host organization for the SOSIG service.) UKOLN and ILRT are located close to each other and have been involved in a number of joint activities, including the EU-funded DESIRE project (DESIRE, 2000a). One deliverable from the DESIRE work was an *Information Officer's Handbook* (DESIRE, 2000b), which describes best practices to support libraries and other organizations interested in setting up large-scale information gateways on the Internet. This handbook, which was jointly authored by staff at ILRT and UKOLN (and others), helped develop both

organizations' expertise and knowledge of quality assurance processes for metadata and, indirectly, led to the work described in this article.

Today, QA Focus is a joint venture between UKOLN and the AHDS (the Arts and Humanities Data Service, based at King's College, London). The change in partnership (which followed ILRT's decision to refocus on their core activities after the first year of the QA Focus project) has strengthened the QA Focus team due to AHDS's additional service responsibilities and experience in a wider range of digitization activities.

The role of QA Focus is to help ensure that project deliverables are interoperable and widely accessible. The remit of the work covers the areas of standards, digitization, the Web, metadata, software, and service deployment. QA Focus seeks to ensure that projects deploy appropriate open standards and best practices in these areas. The approach taken has been published elsewhere (Kelly, Guy, & James, 2003) and is summarized below.

Initial Groundwork Focus group meetings were arranged in the first year, providing an opportunity for QA Focus to inform projects of the service and to gain feedback on work areas that needed to be addressed. The meetings raised the following issues:

- a lack of awareness of recommended open standards in some cases
- difficulties in implementing standards in some cases due to lack of expertise, immaturity of the standards, or poor support for the standards
- concerns over changes in standards during the projects' lifetime

Although it was pleasing to hear that many projects were committed in principle to the JISC's open standards philosophy, it was also clear that implementing open standards would not be easy: projects faced other pressures such as lack of technical expertise, short time scales, investment in existing tools and products, and use of third-party applications and data that sometimes hindered deployment of open standards.

Another activity carried out in the first year was a series of benchmarking surveys of the Web sites provided by the JISC 5/99 projects. The surveys made use of a variety of automated tools, which analyzed the compliance with HTML and CSS standards for the projects' home pages and other features, such as the number of broken links, use of embedded metadata, etc. Although such automated surveys have their limitations (automated accessibility tools need to be supported by manual tests in order to ensure pages are accessible, for example) the surveys were valuable in providing an understanding of common problems and in helping to identify and prioritize areas in which advice was needed.

Briefing Paper The findings of the focus groups and the surveys helped us prioritize the areas in which advice was needed. Since QA Focus was not funded to provide direct support to projects, our advice came in the form of short, focused briefing papers. Currently over seventy briefing papers

have been produced, covering the areas of standards, digitization, the Web, metadata, software, and service deployment.

Advice on Testing Tools There is a clear need for tools to check that resources comply with standards and best practices, including tools such as HTML and CSS validators and link checkers. Although Web developers should be familiar with such tools, our experiences have revealed a number of factors that may result in misleading results:

• Definition of Links: Some links checkers will only check conventional hyperlinks and embedded images. However, links can also be provided using the <LINK> tag for links to external resources such as JavaScript files, CSS files, and metadata resources.
• HTTP Headers: Testing tools should take appropriate actions based on HTTP headers received. Some testing tools report on the output of an HTTP header rather than reporting on the header received.
• Misconfigured Servers: Servers, caches, firewalls, etc. can sometimes be misconfigured, giving misleading findings.
• Personalized Pages: There is an increasing need to be able to test personalized pages. The personalization may be due to a number of factors, including user preferences, browser type and environment, regional factors, etc.

Online Toolkit In order to help embed quality assurance procedures, we have developed an online toolkit that provides a simple checklist. The toolkit helps to focus the developer's mind on key issues and provides advice on the main areas to be addressed. Online toolkits are available in several areas including standards selection, mothballing Web sites, and metadata.

Selection of Standards Although digital library services seek to make use of open standards, there can be dangers in making use of immature standards or not having the resources and expertise needed for the successful implementation of certain standards. We have published a methodology on the selection of standards (Kelly, Dunning, Guy, & Phipps, 2003).

QA Focus Methodology The key deliverable of the QA Focus project has been the development of a lightweight quality assurance methodology. The QA methodology has been informed by the ISO 9000 standard for quality management (International Organization for Standardization, 2004). The methodology requires projects to provide documented *policies* on their technical infrastructure and systematic *procedures* for ensuring they comply with their policies.

Case Studies In order to support the sharing of experiences across the JISC digital library community, QA Focus has also commissioned case studies that provide an opportunity for projects to share their approaches to technical developments. The SOSIG case study illustrates a typical example.

SOSIG Case Study

   The development of a QA framework had its roots at the Social Science Information Gateway (SOSIG). SOSIG (Social Science Information Gateway, 2004) is a well-established Internet resource discovery service for the social sciences, business, and law. SOSIG is based at the Institute for Learning and Research Technology (ILRT), University of Bristol. It is funded by the UK's Economic and Social Research Council (ESRC) and JISC and is part of the RDN. The SOSIG Web site is illustrated in Figure 1.

   SOSIG began life as a pilot project in 1994 and is now considered by many as a pioneer amongst Internet subject gateways in the UK and worldwide. The core of the service, the Internet Catalogue, currently holds over 27,000 structured metadata records (across 17 top-level subject headings and over 1,000 subsections) describing Internet resources relevant to social science learning, teaching, and research. Since its inception, members of the SOSIG team have consistently worked on and developed tools, meth-

Figure 1.

ods, and procedures that support the creation and ongoing maintenance of quality-controlled information gateways. One of the co-authors (Hiom et al., 2003) has been working for the SOSIG service since its launch and has worked closely in the development of the quality assurance procedures for the service. Though always an important consideration, the need for quality assurance procedures has emerged as a real issue with the increasing size and scope of the SOSIG service. To this end, SOSIG now has an established and comprehensive set of procedures that range from the selection of resources to the systematic weeding of the collection. The following case study documents the QA procedures at SOGIG that underpin the creation of high-quality metadata records. These procedures involve subject specialists who are carefully trained according to clear policies, guidelines, and criteria, as well as various automatic checking measures to further standardize the process.

Use of Subject Specialists The records are created and maintained by a geographically dispersed team of over forty subject specialists (known as Section Editors) who select and catalogue Internet resources within a particular subject area. SOSIG relies on a solid set of quality assurance methods that aim to ensure consistency and accuracy amongst the team of specialists. The Section Editors are responsible for seeking out, evaluating, and describing social science Internet resources within their specialized subject area. In addition, the service interoperates with other subject gateway services and therefore aims to ensure that all of its catalogue records are compatible with the wider Resource Discovery Network.

At SOSIG a great deal of time and effort has gone into developing procedures to ensure a consistent approach to the cataloguing process. A thorough training program is backed up with detailed and comprehensive printed and online reference material available to all Section Editors.

Training Each Section Editor receives training on all aspects of working with SOSIG. This begins with an overview of the service from the end-user perspective. An explanation of how the service is used by real people helps to set in context some of the editing procedures—identifying relevant keywords, for example. An end-user perspective is followed by sessions on best practice in locating and evaluating resources and practical training on the online cataloguing center. The workshop is supplemented by documentation in the form of a workbook, as well as a step-by-step guide to cataloguing that includes the following:

1. *The SOSIG Scope Policy,* which outlines the type of resources the Internet Resource Catalogue covers in terms of subject matter, geographical coverage, language, etc.
2. *The Collection Management Policy,* which offers a guide to the selection and deselection criteria for the collection.
3. *The Evaluation Criteria,* which explains how potential resources are evalu-

ated in terms of content, presentation, and any quality assurance procedures that may be in place.

4. *The Cataloguing Rules,* which aim to help SOSIG editors use standard practices when adding records to SOSIG to ensure that records within the database are consistent and of a high quality. The rules include an explanation of each of the metadata fields and how they should be entered (that is, particular formats for dates, names, etc.). The rules also include links to further information such as classification schemes, country codes, etc.

The cataloguing rules document is the most important one in terms of ensuring consistency across the service. With published and publicly available documentation on all areas of the selection and evaluation procedure, Section Editors have a constant resource to turn to while working on the catalogue, while end users can gain a better understanding of what to expect of the service.

Online Tools and Checks SOSIG has integrated a range of online tools and automatic checks (many at the request of Section Editors) into the cataloguing process in an attempt to eliminate errors and inconsistencies prior to the records being added to the catalogue. Controlled vocabularies or thesauri are used for assigning keywords to the records to help in the standardization of spellings, but more importantly, to help users find other related terms and records linked to their topic of interest. Wherever possible, SOSIG uses preformatted authority files to minimize the risk of typing errors creeping into records. Editors are also encouraged to cut and paste URLs into records to avoid errors. Conversely, Editors are encouraged to create freehand, textual descriptions for records. These are seen as an important and value-added aspect of SOSIG. To counter error the system operates a spell-check facility that checks the record as it is submitted to the database and highlights any words it does not recognize. Occasionally this can prove problematic, especially with proper names and technical terms, but SOSIG has included an override function as well as the ability to add particular words to the spell-check dictionary. Online help and access to the cataloguing rules are also provided for Editors through the cataloguing form.

Post-Cataloguing Methods The ideal situation for SOSIG and other digital libraries is to ensure that procedures for quality assurance are robust enough to minimize any editing work after the creation of the catalogue record. Given the volatile nature of information on the Internet, however, it is necessary to implement a number of quality checks on the existing metadata records.

Automatic Confirmation of Record Creation As metadata records are created, an email message is sent to the administrator of the catalogued resource or site to inform them that they have been added to SOSIG and

to give them the opportunity to read the description. Suggestions and amendments can be sent directly to the central administration team for approval. Email content, we have found, is essential for the maintenance of the SOSIG database. We conducted a major one-off "clean-up" exercise in 2003, contacting all administrators of sites that had been catalogued by SOSIG and requesting that they check their record on the SOSIG database for accuracy. This process provided multiple benefits: it not only allowed us to check the accuracy of the records, but it also served as a promotional tool for the service and often resulted in reciprocal linking, suggestions for additional useful material to add to the gateway, and a communication channel for administrators to notify us about major overhauls of their own sites.

Link Checking and Reviewing Given the dynamic nature of the Internet, and the Web in particular, collection development is a major task. Collection management (that is, removing broken links, checking and updating records) at this scale can also be something of a challenge. Many sites often change constantly or even disappear, only to reappear under a new guise. To counter this, an automatic link checker is run over the entire database of URLs on a weekly basis and errors are noted in a report that is made available to Section Editors.

Of course it is not only link errors that need to be considered. Records should also be reviewed on an ongoing basis to ensure that they are still accurate and suitable for inclusion within the catalogue. The Collection Management Policy outlines the principles and process for editing and deleting records. For example, if the information content of the resource has changed so that the resource description and keywords need to be updated, or if the currency or reliability of the resource has lessened over time, the policy has clear directives on how to handle such cases.

Section Editor Workshops Because Section Editors at SOSIG work as a geographically dispersed team, we feel it is important that they are able to get together on a regular basis in order to meet each other and exchange experiences. Consequently, the whole team meets annually to discuss the development of the overall service, to plan changes to their individual sections, or just to brush up on skills generally. Feedback from the workshops suggests that Editors find these events invaluable in that they help to reduce feelings of isolation that can so easily develop within virtual teams.

Summary SOSIG has grown into a large and significant resource. The size of the catalogue raises considerable issues in terms of collection maintenance and the management of a distributed team from many disciplines. The QA tools and procedures described above have developed over a considerable period of time. They are now considered a vital element of the service in that they support the needs of both the central administration of the gateway, the team of distributed Section Editors, and, most importantly, the needs and expectations of the end users.

QUALITY ASSURANCE FOR METADATA

We have given a broad outline of the QA Focus work. We will now focus on the application of this work to the area of metadata.

Purpose of the Metadata

Decisions on the use of metadata in any digital library project should be based on the functionality to be provided by the metadata. The functionality required will influence the metadata standards to be used as well as the architecture for managing and deploying the metadata. Implementing appropriate quality assurance procedures into a project's planning activities and workflow practices will help to ensure that the metadata is and remains fit for its purpose.

Cataloguing Rules

There are a number of problems that can arise for any project using metadata. Probably the most important is the issue of consistency. Ensuring that metadata consistency is maintained is important if interoperability is to be achieved. Where resources are catalogued by more than one person (or indeed organization), the potential for errors in the metadata multiplies. Thus it is vital to ensure that cataloguing rules and a consistent approach are implemented across the board. Services such as SOSIG have adopted a systematic approach to minimize the problems that a geographically distributed service faces when creating metadata. A well-defined interface for inputting metadata, which restricts variation as much as possible, can help this process. Selection lists populated from a controlled vocabulary or fields that only accept data in a particular format are useful ways of restricting variation in metadata creation. The use of authority files will help ensure that naming conventions are followed systematically.

Maintenance

In addition to ensuring that any metadata produced is consistent one must ensure its currency. The evolution of electronic resources is an almost constant activity, and it is important to update the resource's associated metadata alongside the resource itself. Not only will project staff find inconsistencies unhelpful, but machine interfaces will not be able to spot out-of-date information in the way that humans can. The popularity of a resource may fall if users believe it to be out-of-date, even when in reality the resource has been revised recently.

Interoperability

It is important not to be too restrictive when thinking about creating metadata for a project's resources. For metadata to be widely used it must be interoperable. While records may start out only being used in house, ensuring that the project's metadata conforms to standards and maps easily to other metadata schemas will allow the metadata to be used more widely. SOSIG is an excellent example of a pilot project that has evolved

into a service and is now a major Internet gateway used on a national and international scale. Project staff will need to be aware that different cataloguing rules may be used in other environments. As an example, date formats often throw up inconsistencies between the United Kingdom and the United States.

Validation

It is important to ensure that any metadata a project or service produces is validated. If metadata is encoded in XML, it must be validated against a DTD or schema. Metadata creation and management tools should be configured to validate newly created metadata and output it in a controlled format.

Errors may occur in the workflow process: a Microsoft Windows character such as the © symbol could be entered into a database and then embedded in a metadata record in XML format. However, this character is an invalid character in an XML format. The impact of such errors in the record can be considerable: a record that is not spell checked or presented consistently will reduce the impact of your metadata, the service it provides, and its interoperability.

We present a fictitious scenario below in which some of the common problems that can arise when producing metadata have been drawn together.

A multimedia e-journal project is set up. The Dublin Core metadata element set is used to describe published articles. There are documented cataloguing rules in place but, unfortunately, due to a high staff turnover (many staff are on short-term contracts), there are many inconsistencies in the metadata (John Smith & Smith, J.; University of Bath and Bath University; etc.).

The metadata is managed by a home-grown tool. Unfortunately, the author details are output in HTML as DC.Author rather than DC.Creator. In addition, the tool outputs the metadata in XHTML 1.0 format, which is embedded in HTML 4.0 documents.

The metadata is created by hand (with no interface to simplify and control the process) and is not checked. This results in a large number of errors and use of invalid characters (for example, £,—, and &). Consequently, the quality of the records is low.

The metadata describing copyright and access information for the images associated with the articles becomes separated from the images during the workflow process. Since some resources can be freely used by all but others are restricted (used only by the host institution), the separation of the rights metadata from the resources means that the project deliverables cannot be used by third parties.

QA For Metadata Toolkit

We have described a number of areas in which there is a need to address metadata quality when supporting resource discovery. However, metadata can be used to support a wide range of areas, such as maintenance of Web sites, access to e-learning resources, or accessibility. Rather than providing detailed advice for every area in which metadata can be used, we have sought to develop a simple model that can be applied in many areas. Our online toolkit for QA for metadata seeks to ensure that projects have given due consideration to key areas. The QA for metadata toolkit is illustrated in Figure 2.

It should be noted that the toolkit is intended for self-assessment purposes only. A record of the responses is not kept.

The issues addressed in the toolkit are:

- clarification of the purpose for which metadata is being used
- use of an appropriate metadata schema and appropriate cataloguing rules
- appropriate technical architecture for creating and managing the metadata
- procedures for checking the metadata content and syntax
- appropriate training and staff development policies
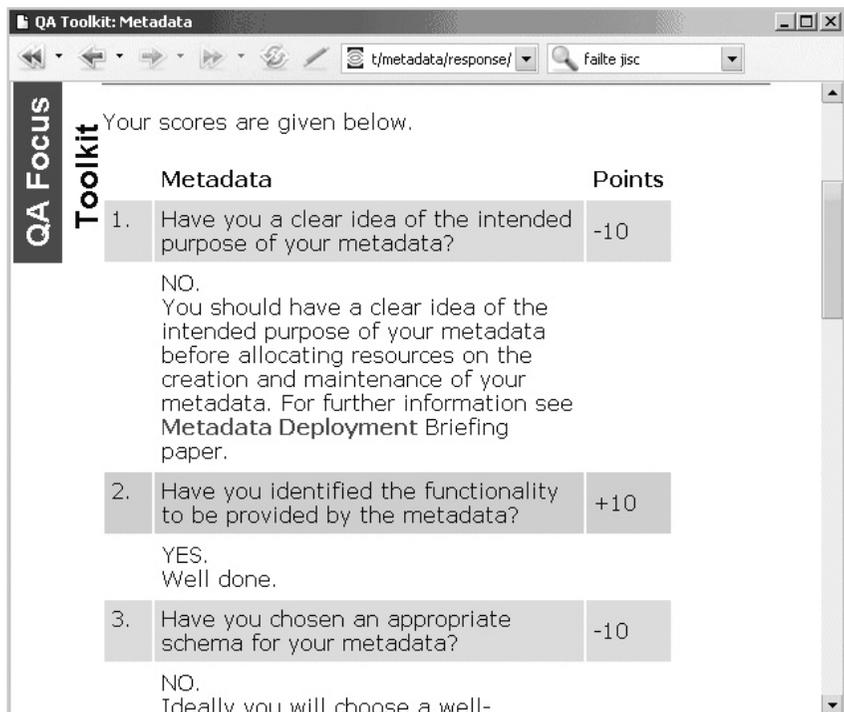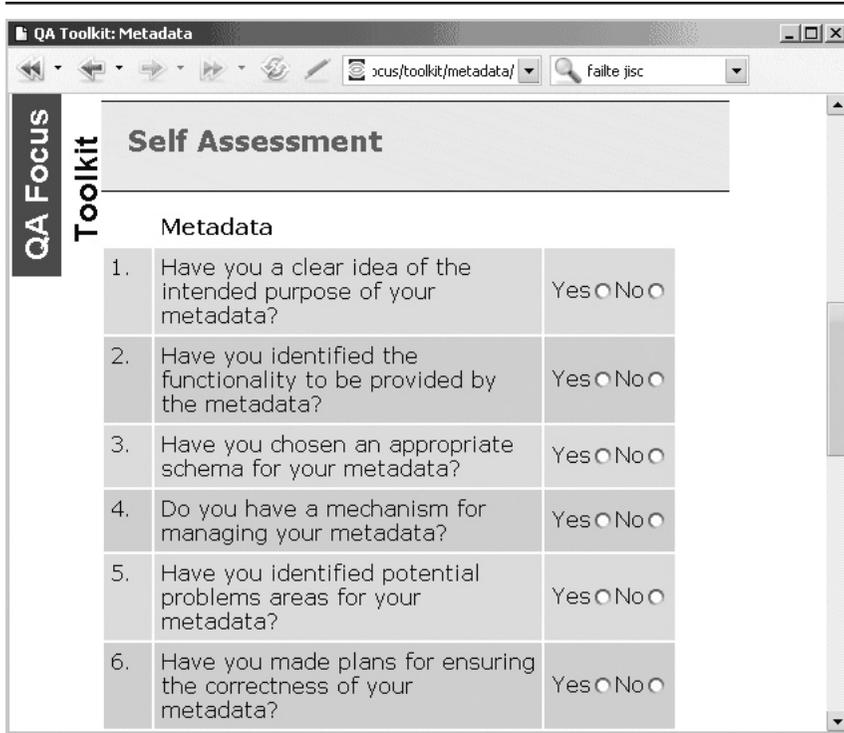- liaison mechanisms with potential remote users of the metadata

We have recommended to the JISC that those JISC-funded projects making significant use of metadata should address these issues as part of the project's reporting procedures. We feel that this lightweight but important approach to the quality assurance of metadata can help minimize interoperability problems and can also be of benefit if a service is to be deployed in a service environment.

Conclusion

In this article we have described the approaches taken by mature subject gateway services such as SOSIG to ensure that they deliver and continue to provide the quality metadata that is essential for an effective subject gateway service. We have sought to generalize this work in the form of a quality assurance framework, which can be deployed by projects and services that wish to make use of metadata. Finally we have described how this quality assurance framework has been extended to support the broad interoperability of JISC's digital library programs.

Metadata is critical to the effective deployment of many digital library environments such as open archives, e-learning environments, and semantic Web applications. Quality assurance procedures will be critical to the effective deployment and interoperability of such services. The authors hope that this article has outlined a quality assurance framework that can be of use to those involved in development work in this area.

Figure 2.

QA Toolkit: Metadata

QA Focus Toolkit

Self Assessment

Metadata

1.	Have you a clear idea of the intended purpose of your metadata?	Yes ○ No ○
2.	Have you identified the functionality to be provided by the metadata?	Yes ○ No ○
3.	Have you chosen an appropriate schema for your metadata?	Yes ○ No ○
4.	Do you have a mechanism for managing your metadata?	Yes ○ No ○
5.	Have you identified potential problems areas for your metadata?	Yes ○ No ○
6.	Have you made plans for ensuring the correctness of your metadata?	Yes ○ No ○

QA Toolkit: Metadata

QA Focus Toolkit

Your scores are given below.

Metadata		Points
1.	Have you a clear idea of the intended purpose of your metadata?	-10

NO.
You should have a clear idea of the intended purpose of your metadata before allocating resources on the creation and maintenance of your metadata. For further information see **Metadata Deployment** Briefing paper.

| 2. | Have you identified the functionality to be provided by the metadata? | +10 |

YES.
Well done.

| 3. | Have you chosen an appropriate schema for your metadata? | -10 |

NO.
Ideally you will choose a well-

REFERENCES

DESIRE. (2000a). *Welcome to the DESIRE [Development of a European Service for Information on Research and Education] project.* Retrieved November 28, 2004, from http://www.desire.org/.

DESIRE. (2000b). *Welcome to the DESIRE information gateways handbook.* Retrieved November 28, 2004, from http://www.desire.org/handbook/.

eLib. (2001). *The Electronic Libraries Programme home page.* Retrieved November 28, 2004, from http://www.ukoln.ac.uk/services/elib/.

Hiom, D., Huxley, L., Place, E., Boyd, D., Cross, P., & Kelly, B. (2003). *Approaches to "spring cleaning" at SOSIG.* Retrieved November 28, 2004, from http://www.ukoln.ac.uk/qa-focus/documents/case-studies/case-study-25/.

International Organization for Standardization. (2004). *ISO 9000 and ISO 14000—In brief.* Retrieved November 28, 2004, from http://www.iso.ch/iso/en/iso9000–14000/index.html.

Joint Information Systems Committee. (2001). *Funding 1/01: National focus for digitisation and quality assurance.* Retrieved November 28, 2004, from http://www.jisc.ac.uk/index.cfm?name=funding_1_01.

Kelly, B., Dunning, A., Guy, M., & Phipps, L. (2003). *Ideology or pragmatism? Open standards and cultural heritage Web sites.* Paper presented at the ichim03 Conference, September 8–12, 2003, Paris, France. Retrieved November 28, 2004, from http://www.ukoln.ac.uk/qa-focus/documents/papers/ichim03/.

Kelly, B., Guy, M., & James, H. (2003). *Developing a quality culture for digital library programmes.* Paper presented at the EUNIS 2003 Conference, July 2–4, 2003, Amsterdam, Holland. Reprinted in *Informatica, 27*(3). Retrieved November 28, 2004, from http://www.ukoln.ac.uk/qa-focus/documents/papers/eunis-2003/.

Social Science Information Gateway. (2004). *SOSIG home page.* Retrieved November 28, 2004, from http://www.sosig.ac.uk/.