USER PASSPORTS AND VISAS

understanding the role of identification metadata

A study conducted by
Book Industry Communication
on behalf of the
British National Bibliography Research Fund

Liz Potter and Mark Bide

Library and Information Commission Research Report 22
British National Bibliographic Research Fund Report 96

March 1999

BOOK INDUSTRY COMMUNICATION

BRG/45

ISBN 1 873671 24 5

ISSN 1466-2949

ISSN 0264-2972

CONTENTS

User Passports and Visas: understanding the role of identification metadata

By Liz Potter and Mark Bide [1]

1 Executive Summary

This paper surveys current practice in the identification, authentication and authorization of users of electronic resources in the UK. We have focused on access to academic journals, researching the ways in which access to those journals is managed both by their academic and by their commercial user communities.

At present, access management in this field in the UK is largely managed via username and password credentials or IP validation, or a combination of the two. We consider the principles on which these approaches operate, and draw out the ways in which they fall short of the requirements of various players in the chain. In particular, users are far from achieving the seamless access to resources which must be the goal; information support staff bear an enormous administrative burden; and there are significant inefficiencies caused by the lack of standardization in the data used to support these processes though the chain.

We discuss associated issues in the management of resources in the digital world, including identifiers in use for the objects that users access and the standardized definition of the terms of availability under which those objects are used. We also address the critical issue of user privacy, which arises as a result of the detailed usage data that management systems are now able to collect.

From our survey, we evaluate the potential for managing these processes more effectively in a different way: by employing User Passports and Visas. The model proposes that each individual possesses a User Passport which identifies both who one is and something about the groups and classes to which one belongs. Within the passport, Visas exist which indicate that the individual has certain rights in respect of certain resources. Some of these rights may derive from institutional membership; others might be acquired via an individually negotiated arrangement.

We discuss the processes required to make the model work in practice, if it is to ameliorate the deficiencies of current practice and encourage standardization throughout the information chain. We conclude that the model could offer advantages over current practice, provided that significant modifications were made in current behaviour and practice.

Finally, the study considers what data would be required to provide adequate proof of identity, status and institutional membership, and concludes that there is a great deal of work to be done in standardizing the expression of this metadata.

[1] Mark Bide & Associates, London office: 105a Euston Street, London NW1 2EW: liz@markbide.co.uk. We have been valuably assisted in the development of this paper by Sally Morris, Secretary General of the Association of Learned, Professional and Society Publishers, sally@morris-assocs.demon.co.uk

2 Introduction

This paper is a successor to two recent papers from Book Industry Communication. Their aim, broadly, was to provide an overview of the standards required to support certain facets of transactions in "content" on networks, particularly the World Wide Web. The first paper – *Unique Identifiers: a brief introduction*[2] – focused on standards for the unambiguous identification of specific items of "content" in such a transaction. The second – *User Identification and Authentication: a brief introduction*[3] – concentrated on the identification of users – that is, those who access the content once the financial transaction has been made.[4]

This paper follows particularly closely on the second paper, while assuming a knowledge of the standards issues described in the first. It is a study of the feasibility of implementing one model of user identification, authentication and authorization, that of "User Passports". This model[5] proposes that, as in the physical world, a digital passport should identify a person and offer proof – "authentication" – of that identity (as a photograph authenticates an identity in a physical passport). Further, it should record the content that that person is authorized to access, rather in the way that Visas in passports record the country or countries to which a person is authorized to travel.

If it were to be implemented, this model of digital passport and visa would require an underpinning of standards, such that users, and the rights they have acquired to access particular resources, are described in a standard fashion. This paper considers the feasibility of its implementation in two stages. First, we look at the ways in which identification, authentication and authorization are currently handled for academic journals and full text databases, since this is the area in which these issues are already being faced on a day to day basis in the UK. This may be familiar ground, but it provides a 'reality check' of our current practice, and indicates how that practice may be falling short of the requirements of users.

Second, we assess the requirements for the identification of users and their rights in the passport and visa model and the role that "identification metadata" would play in its implementation.

Intellectual effort has been devoted elsewhere to many of the issues which surround this topic. Initially, then, we offer definitions of the terms we are using in this paper, in order to locate ourselves in the debate and to avoid potential confusion. We must also make clear those areas of the debate that this study does *not* aim to address.

2.1 Definitions

Inevitably, and deliberately, our definition of terms owes much to the previous literature on the topic. We are particularly indebted to the White Paper issued earlier this year by the Coalition for

---

[2] Green B and Bide M *Unique Identifiers: a brief introduction* (September 1996, revised March 1997) London: Book Industry Communication. Available at http://www.bic.org.uk/bic/uniquid.html This paper focused on standards in use and in development at the time of writing. Developments in standards that have taken place since March 1997 – most notably in the Digital Object Identifier (DOI) – can best be discovered through the International DOI Foundation web site at htttp://www.doi.org

[3] Bide M and Hing T *User Identification and Authentication: a brief introduction* (February 1998). London: Book Industry Communication. Available at http://www.bic.org.uk/bic/userid.pdf

[4] We recognise that 'end users' may or may not have been the purchasers in the financial transaction.

[5] Originally proposed by Sally Morris. See Morris M "Standards for Electronic Rights" SISAC News (ISSN:0885-3959), Vol 13 no. 1, Summer/Fall 1998, pages 8-13. See also Bide and Hing 1998, p12.

Networked Information,[6] which we have found consistently useful and thought-provoking in our study.

2.1.1 People and objects

In this paper, a **user** is a person making use of a resource on a network. A **resource** is the generic term we shall use to describe "objects" or "content" made available to users on networks; a resource may be a document, a series of documents, a web site, a database, a number of databases. A **user community** is a community that has established a right to use a resource, usually via a licence; the institution which negotiated this licence on their behalf is referred to as the **licensee institution**. An institution might be a commercial organization, a society, an educational establishment.

We use the term **resource operator** to refer to those who offer users access to resources on the network. This term covers both publishers, in those cases in which they manage access to their resources themselves, and intermediaries, who offer a point of aggregated access to material from various sources. **Resource providers** are the original providers of the resource at a more basic level – those who compile and publish the resource, and from whom typically access to the resource is licensed. These providers may also offer a point of access to their resources, or may offer access via an intermediary operator, or both. We use the two separate terms to distinguish between the *functions* they are performing.

2.1.2 Processes

The **identification** process represents the user's claim to an identity. **Authentication** is the process in which the user offers proof that he or she has a right to that identity. In an institutional context, the user is effectively proving that he or she really is a member of a particular licensee institution. That proof must take a form agreed between the licensee institution and resource operator prior to the user's attempt to access the resource, a form that the resource operator can understand, check and validate.

This context is the one in which most users have access to academic journals – that is, most individuals currently use academic resources by virtue of belonging to a subscribing (licensee) institution.[7] However, we also wish to consider the identification of individuals who do not belong to institutions.

If authenticated, a user has established a right to assume an identity, but has not, by this process alone, identified herself as a particular individual in the physical world. A single user might, indeed, legitimately have a right to more than one "identity". This is an issue to which we shall return.

Authorization is the process in which a resource operator determines what the holder of a particular identity is allowed to do – for example, whether that identity is permitted to access a resource. **Access management** is a term that is often used to cover both authentication and authorization.

---

[6] Lynch C (ed) *A White Paper on Authentication and Access Management Issues in Cross-organizational Use of Networked Information Resources*. Coalition for Networked Information. Available at http://www.cni.org/projects/authentication/authentication-wp.html

[7] We place our emphasis on academic journals in view of the scope of the study. However, it does not seem to us that the discussion below is solely applicable to academic journals or even academic resources.

2.2 Scope

This study has three main aims:

- To provide a survey of existing systems and approaches to access management, primarily in the area of academic journals (see Section 3);
- To define, at an abstract level, the processes which would be required to manage a "passport and visa" system for identifying users and their access rights (see Section 4.2);
- To consider the definition of a data dictionary and standard metadata set to describe both users and rights they have acquired (see Section 4.3).

This study, then, does not constitute an overview of the full range of ways in which access management can be handled and an assessment of their advantages and disadvantages. Nor does the paper address the technology infrastructure which would be necessary for the implementation of the model. These are critical issues, of course, but they are out of scope for this study. Again, we refer readers to the second of the BIC papers mentioned above and to the CNI White Paper for further discussion.

3 Current practice

In determining what is current practice, we focused primarily on access via the Internet to academic journals and databases in the UK, and spoke to a range of Higher Education Institutions (HEIs), Commercial Organizations, Data Services Providers, Intermediaries and Publishers.

Access management in this field is largely being managed via username / password credentials or IP validation, or a combination of the two; cookies are also used, but this is less typical. The concerns of user communities – be they commercial organizations, educational institutions or societies – were found to be largely equivalent, rather than differing according to type of institution.

We begin by charting current practice, and the principles on which it operates (Section 3.1), and draw out some of its implications in the sections which follow (Sections 3.2 – 3.7).

3.1 Current practice in access management

3.1.1 Usernames and passwords

This approach to access management asks the user to respond to a 'challenge' when he or she attempts to access a resource. The user types in a username as a claim to an identity, and a password as proof of that identity. Passwords constitute 'proof' on the basis that (in theory at least) they are confidential to the user, and cannot be given by anyone other than the user who possesses the partner username.

Users are currently obliged to manage multiple usernames and passwords. The user of a library, for instance, might access a number of the library's internal resources by using one username and password pair, or one username and a variety of partner passwords. These username and password strings are allocated by the institution, usually when users have enrolled (in the case of students) or have been added to the payroll (in the case of staff). Usernames typically bear some relation to the name of individuals in the physical world – for example, they might be based on surname and initial, with truncation and adjustment for uniqueness – but the relationship is neither formalized nor standard within or between institutions: one cannot derive a username from a personal name with certainty.

In addition, that user will be allocated other username / password pairs to permit access to a range of external resources – that is, resources operated outside the institution. In these cases, resource operators distribute a 'master' username and password to a site administrator at a subscribing institution. The administrator then creates sub-groups of users, allocating them each a username / password pair in an appropriate hierarchical structure. If the system allows, administrators generally try to use the same username as standard, coupled with a different password. This is not always possible.

In an institutional context, the resource operator can derive the institution to which a user is associated from his or her username and password. They must then map these to individual access rights – that is, they must validate a user's claim to a right of access to a particular resource by verifying that the resource provider's subscription information for the requested content includes the user's institution (or the individual user, in a small minority of cases).

3.1.2 IP validation

This method of access control requires the licensee institution to guarantee to the resource operator that all resource requests deriving from a range of IP addresses[8] will be legitimate requests. To authorize access, the resource operator simply checks the IP address against the information supplied by the licensing resource provider.

Some resource operators use IP validation on its own, and one can see its attractions: it is relatively simple in administrative terms, and for the user it has the advantage of providing what appears to be 'seamless' access to resources. Others use IP validation in addition to username and password control, in order to exert a 'double check' on authenticity. In cases where users access content through another third party, the demand for another username and password is so irritating for the user that it is increasingly regarded as unacceptable. Seamless access is controlled in these cases by confirming 'behind the scenes' that the user is making his or her request from a machine with a recognized IP address.

3.1.3 Cookies

Access management can also be handled by using "cookies", small machine readable files installed on the user's machine. These provide a mechanism for the resource operator to provide access to a resource for the life of a subscription. They also provide the opportunity to store information about the way the resource has been used – such as how many times the user has accessed the server and when, and what searches he or she performed. From the resource operator's point of view, then, the cookie constitutes 'proof' that this is a valid access request. From the user's point of view, once the cookie is installed, he or she need never key any credentials, providing 'seamless' navigation. A cookie also provides the basis for customization of the interface for the individual user, which is widely seen as providing much of the real potential for commercial exploitation of the World Wide Web.

The widely recognized draw-backs of handling access management by using cookies are dealt with in Section 3.2.3.

3.1.4 Athens

The Athens access management system[9] properly deserves mention here, as a particularly widely used system to support username and password authentication. It is in use in all UK HE institutions to enable controlled access to subscription services provided by publishers and data suppliers and made available via the Data Service Providers BIDS, EDINA, MIDAS and NISS. Other resource providers – both in the educational and commercial spheres – are investigating the feasibility of using Athens, but as yet it has not been implemented elsewhere.

To access these resources, each user must have an Athens account. The management of accounts is distributed to site administrators, enabling each institution to manage its own users on an appropriate hierarchical model. A 'domain administrator' – usually assisted by other administrators with responsibility for particular parts of the institution, such as departments –

[8] An IP address uniquely identifies a particular computer on the Internet. IP addresses are 32-bit binary numbers which are given in four-part decimal numbers, each part representing 8 bits of the 32-bit address (for example, 123.456.7.891).

[9] See www.athens.ac.uk

manages user accounts and groups, and the authorization properties that are inherited via those groups.[10]

It operates on the principle of single sign-on – the same username and password allows access to all permitted resources. However, from the user's point of view this means single sign-on *for one set of required resources only* (ie, those made available via the four resource operators listed above). For resources made available via other operators, users must sign on again, using different identification and authentication data – that is, a different 'identity'.

3.2 Implications for users

3.2.1 Usernames and passwords

The proliferation of electronic resources has seen a corresponding increase in the number of usernames and passwords that users are required to remember. One librarian we spoke to said it was not uncommon for a user to be given as many as a dozen username/password pairs when joining the institution in order to access the various resources he or she would need. This is clearly an untenable situation: users cannot be expected to remember an ever increasing number of passwords. Indeed, in order to remember them, they are certain to have to write them down, immediately compromising their security (and therefore their value in access control).

The situation is bewildering not only for end users, but for information support staff as well – 'intermediary' users of the systems.[11] Since it tends to fall to them to allocate passwords, they bear an enormous administrative burden, allocating new usernames and passwords when users forget or lose theirs, and deleting the previous ones from their management systems. Password replacement is one of the biggest irritations for almost any librarian working in an institution which provides access to networked resources. Bigger institutions are finding it almost impossible to administer the password allocation process. Indeed, a number of librarians working in large corporate R&D centres told us that they would not alert their users to the existence of e-journals if they required password validation.

3.2.2 Is IP validation the answer?

Pressures such as these are causing an increasing number of the resource providers we spoke to in our study to move away from their current username and password control and offer subscribers authentication via IP validation, or at least to offer it as an alternative. This will reduce the burden on library or computer services staff, and will meet the criterion of seamless access from the user's point of view.

However, access management using IP validation is not without its problems. Critically, IP validation will only allow users to access resources when they are using machines located in the institution, whose IP addresses fall within the recognized range. This most frequently causes problems when users wish to work from home or elsewhere. To meet this need, resource

---

[10] The model depends, to some extent at least, on "intelligent" identifiers. The arguments for and against intelligence in identifiers has been substantially rehearsed elsewhere.

[11] See, for example, Hamaker C "Chaos – Journals Electronic Style" *Against The Grain* Dec 97 – Jan 98. It is telling that many library projects recognise the frustration caused by the multiplication and repeated entry of user IDs and passwords in their terms of reference. For example, the Decomate project at the British Library of Political & Economic Science at the London School of Economics takes as its starting point that the principle of single user sign-on is essential. Of particular interest here is the Candle Athens Integration Project, which is attempting to harmonise local access to networked information services at South Bank University with access to resources using the Athens authentication service (see http://www.jtap.ac.uk/projects/jtap-627.html).

operators who use IP validation also allocate a number of usernames and passwords to the institution, which are handed to individuals for use at home, on sabbatical, on temporary transfer. Again, the resource operator devolves responsibility to the institution for ensuring that these passwords are only given to users covered by the licence terms. (It is worth noting again that each resource operator will provide a separate set of username / password pairs, relating only to their own resource(s), multiplying the number of pairs the remote user is required to handle.)

Resource providers relying on IP validation also tend to fall back on usernames and passwords if users are accessing resources via proxy servers. In addition, IP validation is insufficient when financial transactions with individuals are involved, for it does not identify at the individual level, merely at machine level. One publisher we spoke to, who finds IP validation satisfactory for institutional access to its journals (using username and password for home access), recognizes that this won't meet the need for a future service where users will be expected to pay individually (by credit card pre-payment in to an account) for different levels of access.

3.2.3 Are cookies the answer?

Cookies are used by some publishers to cover those cases, such as home working, in which they cannot use IP validation. They have reservations about using usernames and passwords, concerned that they will be distributed by users to other 'illegal' users – circulated on email lists, for example.

However, there are drawbacks. For one thing, cookies identify workstations rather than individuals; this means that users must install cookies on each and every machine from which access is required, and that resource providers know little about their *users*, but rather about their *machines* (as is the case with IP). Further, some users are reluctant to accept the installation of cookies on their machine, both for security and privacy reasons. These are significant concerns, and should be borne in mind when considering the implementation of passports and visas, which have something in common with the cookie approach (see further Section 4.2).

3.2.4 Keeping information about access rights up-to-date

Rights to access change frequently, and must be reflected in management systems.

On the one hand, the constituency of legitimate users changes *within* institutions. There are two facets to these changes. First, there is the changing institutional population as a whole – to take the most obvious example in the academic community, there is an annual cycle of summer graduates and autumn freshers whose access accounts must respectively be closed and opened. Second, there is the changing membership of particular subgroups, such as courses, classes, research groups.

The access management systems currently in place, including the Athens system, do not respond automatically to such local changes. Changes in institutional data imply a change in access rights, but this change is not automatically inherited by the user's account. Batch processing is used to update rights information where possible – for example, at the end and beginning of an academic year. This helps institutions to deal reasonably efficiently with our first type of institutional change, that of the changing institutional population as a whole.

However, the second type of change, that of the changing membership of particular subgroups, is a greater administrative burden. When a student changes course options, for instance, and therefore loses certain access rights and acquires others, manual intervention is required at the level of the personal user account to reflect the change which now exists on the university system. In practice, users try to access resources, find they unable to, and approach librarians to discover the root of the problem, which lies in the fact that their account has not been updated. This is

what might be called an “aggravation-based” approach to updating, an irritation to both end user and ‘intermediary user’ (information support staff). This is perhaps an area that could be addressed in future system developments.

On the other hand, the rights of institutions *themselves* are subject to change – if they cancel a subscription, for instance. In this case, the resource provider must inform the resource operator(s) of the relevant change. Currently, this information is being passed to resource operators in non-standard ways, which they frequently re-format to suit their requirements. Again, we see a lack of standardization in the chain, which reduces the possibility of effective interoperability between systems. If the use of electronic resources increases, as it surely must, it is imperative that we develop standards for transmitting such data that will enable these processes to be machine-moderated, as far as possible.

3.3 Implications for identification of the individual

Clearly, then, there are significant drawbacks both with username and password systems, and with IP validation. Most access management systems are currently adopting, or planning to adopt, multiple approaches, because of the deficiencies of each on their own – deficiencies from the point of view of everyone in the chain. We turn now to consider how current practice is handling identification of the individual and what the implications of this current practice may be.

3.3.1 Difficulties for individuals

Access to academic electronic journals currently tends to replicate the manner of access to printed journals – that is, access operates primarily on an institutional subscription basis. Publishers, aggregators and users have had to come to terms with significant technological changes, but the underlying subscription model has remained predominant, and may well continue to do so in the case of academic journals. However, in other areas – for example, where individual subscribers are not uncommon, or are even the norm – this model is clearly impracticable.

Where users are not allied to institutions, the username and password system becomes significantly more cumbersome – difficult for the user, and with administrative overheads incommensurate with return for the provider. The publisher must alert the resource operator that there is a (‘non-institutional’) user in, say, Madrid, the resource provider must contact them with a username and password and then create their access rights in the format required by their authorization system. In other cases, individual users are simply refused access if they are not part of an institution.

3.3.2 What are we identifying?

In most of the systems currently in use, individuals are identified purely as members of a subscribing body – whether a society, a company, a university, a faculty within a university.

When a user accesses a resource, the provider simply checks that he or she is a member of an institution with a valid subscription, on the basis of their username and password and/or IP address. A significant corollary of this is that the resource provider knows nothing about any other rights the user may have acquired – by being a member of another society, or an alumnus of an institution, or simply by having paid for an individual subscription. As a result, users must offer a different ‘identity’ every time they wish to access a resource to which they have access by virtue of being a member of a different ‘group’ or ‘class’.

Further, the nature and allocation of usernames raise an important implication for our purposes: when a user types in a username, they are simply claiming the right to use an allocated name in a particular domain. This identity is allowed to perform only a very limited range of actions –use

the library's bibliographic services and email, for instance, or use one set of journals aggregated by a particular resource provider.

Names or identities, then, are authorized by different organizations, from which derive different access rights; rights to a range of resources are not assigned as attributes of a master name. It is quite possible that, as the CNI White Paper contends, the sort of centralized identity system which would support the attribution of rights to master names would represent "an unacceptable concentration of power", as well as being "technically impractical" at the scale required. However, this should not blind us to the practical implications and inconveniences of our current systems.

3.3.3 Whom are we trusting?

The locus of authorization in the institutional model lies squarely with the institution. The resource operator releases the resource to institutional members on the basis of a trust relationship with a representative – usually a librarian – of that institution. This is the case in the Athens model, for instance: responsibility lies with institutions – specifically, with their administrators – to manage their groups in accordance with licence terms.

Gaining access to resources is cumbersome, if not impossible, for individuals precisely because there is no one to vouch for them – they are not allied to an institution who will 'authorize' them. Any model of identifying users will need to address the question of who vouches for the veracity of an individual's claim.[12]

3.4 What are you looking at?

We have so far talked about 'access rights' without being specific about *what* users have the right to access – that is, how the object being accessed is defined. At present, users are generally pointed to a URL which relates to the journal in which the article they require is published, which they can access by 'drilling down' through the hierarchy. Some resource providers are using Infobike IDs[13] to identify at the article level. Identifiers such as the SICI, ISSN and DOI are not yet in general use for the purpose of identifying objects, although many publishers and aggregators await developments in the DOI with interest, and expect to be implementing it soon.

It is important to note that the inability to move directly to an article from a bibliographic reference, without navigating through a journal hierarchy, was highlighted as a significant frustration to users in the course of our research. Resource operators might be well advised to address this problem.

3.5 What are you doing?

One of the more significant gaps in the framework for the management of resources in the digital world remains our inability properly to express "terms of availability".[14] Essentially, the only right which we are currently capable of defining in electronic terms (as is implicit in this paper) is

[12] It is possible that certificates and "trusted third parties" might offer a way forward here. For a detailed discussion of the technology, see Bide M and Hing T (1998).

[13] Developed as part of the eLib project Infobike. For details on the project, see http://www.bids.ac.uk/elib/infobike/homepage.html

[14] On this issue, see further Martin D & Bide M (1997) *Descriptive Standards for Serials Metadata and Standards for Terms of Availability* Two related eLib Supporting Studies commissioned by UKOLN. Available at http://www.ukoln.ac.uk/models/studies

the right of "access". We do not believe that there is currently an adequately well formed intellectual framework for defining terms of availability other than "access" – for example, defining rights in respect of viewing, or saving, or printing.[15] This brings us on to an important issue, the question of "enforcement" of terms of availability. It seems to us that there is every possibility that there will be a major dichotomy in the delivery of networked intellectual property, between models which depend primarily on "compliance" to protect rights owners and those which depend entirely on systematic enforcement ("trusted systems"). At the level of the individual user, questions of identity are clearly less important in the case of trusted systems since the user is technically unable to "do" anything with the content that has not been authorized by the rights owner. It is probably largely in the compliance model that there is a real need to ensure that an individual has the rights to which they lay claim.[16]

It may well be that, within the environment that we are studying in this paper – the delivery of information to academic and business organizations – that the only "right" which is really significant is the right of "access", with constraints on abuse of that access imposed by a compliance culture rather than by technology.[17] The price to be paid for this may well be that the organization will have to take rather more responsibility for its members than might otherwise be the case. This is an issue that is being robustly debated between academic rights owners and librarians, in relation to CLA licensing in the area of retrodigitization rights.[18]

3.6 Usage data

'Usage' or 'management' data tracks the pattern of use of resources. It can be faceted by user – this user accessed so many resources – or by resource – this object or service is being used this frequently. Publishers can use this data to make a case for re-subscription to customers, and, of course, it is useful for librarians in a similar fashion.[19]

In general, usage data is currently being distributed at summary level (although resource operators do store more detailed data – at the individual user level – if it is technically feasible).[20] It is typical for resource operators to pass usage data to publishers at an institutional level – for example, an academic institution in Illinois accessed a certain number of articles (not identifying which specific articles, however) in a particular journal in a certain time period. Alternatively, the resource operator might identify a specific article, but give less information at the level of the user community.

---

[15] Although considerable effort has been put into developing language – vocabulary and syntax – to describe those rights, most notably by Mark Stefik and the team at Xerox who have developed DPRL. The need for additional work in the whole area of rights definition is now recognised by those working on the INDECS project.

[16] Trusted system models seem to be tightly bound to hardware identification rather than user identification.

[17] In general terms we believe this to be desirable. Any technical enforcement system will add overhead in terms of both cost and technology. There is no advantage in adding to costs simply for the sake of doing so, if broadly the same result can be achieved in other ways.

[18] The argument that rights owners must take action themselves against individual infringers of copyright in the digital environment is widely regarded – by rights owners at least – as untenable. The price to be paid for the adoption of a compliance rather than a technologically controlled culture in the provision of resources to HE may prove to be the acceptance by HEIs of a much greater degree of responsibility for the actions of their members.

[19] Although we have recently heard – anecdotally – that some resource operators are unwilling to share any usage information with their customers; this appears to us to be ultimately a self-defeating policy.

[20] See below, section 3.7, on the privacy issues which are the basis of users' and institutions' concerns about the transmission of data at more detailed levels.

We should recognize the poverty of some of this data, however. What does it really mean to know that a certain institution made so many 'hits' on a particular journal?[21] It does not tell us anything about what they actually *did* – whether or not the users even read the material, for example. Or again, what does a single download figure mean? It certainly does not reveal how many copies the user might have made of the output. There is an analogy in library borrowing data – ten readers might have borrowed a book, but that tells the library little about what they did with it other than removing it from the premises. We are generating an enormous amount of *quantitative* data, but should perhaps be wary of its *quality* (and therefore its value).

3.7 Privacy

Statistics about the use made of resources by individuals leads closely to the issue of privacy. The extent to which usage data undermines privacy depends on the amount of contextual information about the user that is available to its interpreter. Following the CNI White Paper, we should distinguish between different levels of information revealed about the user. To remain totally **anonymous**, an access management system must make it impossible for repeat access to be identified. If repeat use is recognized, but specific user identity is not – for example, if the resource operator simply knows that 'User X' has been accessing the resource in a certain pattern – we have what the White Paper calls **pseudonymous** access. If repeat use is recognized, and supplemented by demographics, but actual user identity is still not revealed, the system is providing **pseudonymous access with demographics**. Alternatively, **actual identities** may be revealed, which may or may not be **supplemented with demographics**.

Each of these approaches clearly compromises privacy to a different extent.[22] Although usage data is typically managed only at institutional level in our area of study at present, as we outlined above, we should be aware of the options that will become available as access management systems mature.

There are important moral, political and economic stances to be taken here. For our part, we stand by the position stated in the previous Book Industry Communication paper: we believe that, in the absence of compelling arguments to the contrary, the privacy of the individual should be regarded as paramount. In addition, as we stated there, there are compelling reasons for maintaining privacy at institutional levels, too – many corporations fear leaving a 'trail' of evidence about the nature of their developmental research. Indeed, we understand that it is quite usual for researchers – both academic and corporate – to order an array of 'spoof' articles and journals on a regular basis, in order to put potential 'spies' off the scent. This illustrates quite how seriously privacy issues are taken, in commercial, academic and personal contexts. It also suggests that, at present, we are forced to adopt some rather wasteful ways of maintaining our privacy.

[21] Indeed, there is no agreement about what is being counted, and what actually constitutes a 'hit'.

[22] Of course, we should also recognise that, in addition to the usage data revealed as a by-product of an access management system, users may also choose to reveal more about themselves – for example, in order to gain access to further services.

4 The role of identification metadata

The increasing availability and increasing use of electronic resources will require both providers and users of information to come to terms with the deficiencies that persist in current practice. If we are to develop access management appropriately, we must develop standard ways of identifying every element of the transaction – the end user, the object, the rights the user acquires in relation to the object, the use made of the object, the vendor, the rights owner or owners. This would provide real advantages for everybody in the information chain, if only because it would become possible to implement business arrangements more effectively (and therefore less expensively). Users need to be able to enjoy easy and seamless access to a range of network resources, preferably both within and without the institutional context.

Here we are particularly concerned about identifying *users*, but much work continues to be required in the standardised identification of every other element of the transaction. We are optimistic that considerable forward movement is being generated both by the International DOI Foundation and others – although anyone who follows the correspondence on issues relating to the identification of resources will know just how complex an issue this is proving to be. However, there are good commercial reasons why we can expect to see some issues resolved in the relatively short term.

Similarly, as we write, work is commencing on a major European-funded project, INDECS;[23] the primary objective of this project is to develop "practical, fast-track solutions to some basic e-commerce infrastructure issues" through the "practical interoperability of digital content identification systems and their related rights metadata within multimedia e-commerce". This project, which involves representatives of the creative industries from across the media, is primarily seeking to develop metadata standards for business-to-business transactions. Nevertheless, we would equally expect it to have a major impact on transactions with end users insofar as it is seeking to describe the fundamental building blocks needed for the management of all dealings in intellectual property in the electronic environment.

In this paper, we seek to encourage the development of a similar framework for the identification and authentication of users. Such a framework should be entirely independent of the technology employed to implement it.

4.1 The passport and visa model

It will be useful at this stage to summarize the fundamentals of the passport and visa model.

The model proposes that each individual possesses a User Passport which identifies both who you are and something about the groups and classes to which you belong. By 'groups' we mean educational institutions, business communities, societies. By 'classes' we mean groups whose membership is not directly (or in some cases, not at all) tied to a formal institution, such as student, under-18, or even nationality.

Within the passport, Visas exist which indicate that you have certain rights in respect of certain resources. Some of these rights may derive from institutional membership – for instance, a library to which you belong subscribes to a particular electronic journal, which means that you can

[23] Interoperability of Data in E-commerce Systems.

access it (however inadequate a definition "access" might be[24]). Others you might acquire via an individually negotiated arrangement, such as making a personal payment for a particular service.

The theoretical model itself implies nothing about where passports are physically located; or who allocates visas; or the technology required to deliver the model. We offer below some thoughts about the processes required to make the model work in practice. Our contention is that the model "works" if and when it ameliorates the deficiencies of current practice and encourages standardization throughout the information chain.

4.2 Towards a working model

4.2.1 Visas on workstations

Clearly, users must be able to offer their passports – incorporating their visas – for inspection from every workstation from which they wish to access networked resources. This implies that they are deposited on every hard disk they wish to use, or that they are available on smart cards or some similar portable medium.[25]

This implementation of the model starts to look rather too like a "super-cookie", readable by any resource operator, with all the implications that that might have for user anonymity. However, within the context of institutional membership,[26] it is possible that an alternative could be found.

4.2.2 Trusting the institution

The institution "knows" the identity of its members. The institution also "knows" what rights it has negotiated on behalf of specific classes of its members. The individual user could be required to sign on to an identification/authentication/authorization service at her institution at the start of every online "session". The institution, "knowing" what rights were associated with that individual, could ensure that an indication of those rights were placed, appropriately encoded, on the workstation that that individual happened to be using, wherever that workstation happened to be. This authorization could easily be time limited (in the same way as a cookie, but maybe for a shorter time) and could be encrypted in the form of a digital certificate authenticated by the institution.[27]

Unlike the typical "cookie", this certificate need not identify the individual user, simply the status of that user and the rights which this status confers on her. The data would need to be readable by all relevant resource operators, but would need to be structured in ways which protected the commercial interests of resource providers (so that it was not possible for a user, visiting a particular resource, incidentally to pass information to that resource about any rights they might have to access another resource).

Such an approach, it seems to us, could have considerable value:

---

[24] See 3.5.

[25] Our reservations about the use of smart cards for the implementation of this model is simply their lack of ubiquity. It is clear that they could have considerable application in managing authentication and authorization (although also, presumably, easily misused). It is clear that we should continue to watch the development and adoption of smart card technology on the basis that it could provide a very useful platform for the implementation of the passport and visa model. We do not seek, though, in this report to consider issues of specific technological implementations at any great depth.

[26] In its broadest sense – educational institution, business, scientific society.

[27] See Bide & Hing 1998.

It would involve the user in a single act of identification and authentication in every session when they were accessing resources in general (every day, perhaps) but would not depend on location. In other words, some of the drawbacks of both the username / password and IP validation methods would be avoided.

It would protect the identity of the individual user from being identified with access to any specific resource, both by the resource provider and by the institution, while providing (potentially at least) "pseudonymous access with demographics" (which is important to both the provider and the institution in assessing the way in which a resource is used, and hence its value).

It would greatly simplify the role of resource operators in the chain, particularly where these were intermediaries, since verification of the claims of an individual both to a specific identity and to a specific set of rights would be unnecessary. This is a critical point, and would certainly please the resource operators we spoke with in our study, for whom the mapping of a claimed identity to rights information furnished by the provider is a significant burden.

It would be a relatively low-cost approach, since it would involve the adoption neither of the relatively high technological burden of "trusted systems" nor the requirement for "trusted third parties" to be involved in every transaction (particularly since all transactions would leave some sort of audit trail).

However, there are some other issues to be considered:

It would imply the development of a degree of trust between institution, resource operator and resource provider which is often singularly lacking in the current model. The model relies on everyone in the chain placing absolute trust in the user's visas, and thus in the institution to which he or she belongs, who had granted the visas. This would require adherence to high standards of probity; it would also place on the institutions a much greater responsibility for failures of compliance.[28] It would also demand very tight adherence to data standards, in the sense that the data held in the passport and visa would need to be readily and unambiguously comprehensible in use.

It would not provide the complete anonymity which is typically required by commercial organizations – although this is not insoluble, it would require an even greater degree of trust since individual transactions would have to become unauditable.

The model only works for rights which the institution itself acquired on the user's behalf, and does not cater for their other rights. It relies on the placing of absolute trust in institutions to vouch for their members and their rights, but this does not help individuals who have rights which are not derived from institutional membership. Who vouches for them? Without the existence of some sort of Universal Certificate Authority – or more plausibly a number of Certification Authorities[29] – with the power to grant such visas it is hard to see how this model can cope with individuals. It remains to be seen whether such centralization of authority will be considered acceptable and whether a commercial model can be developed which would allow the economic use of such services to manage authorization of access to resources.

There is a direct analogy with physical passports here: governments vouch for their citizens, and one cannot obtain a passport if one is stateless. We are doubtful that Government would see itself

[28] See also footnote 18.

[29] The Post Office, for example, may be setting up as a Certification Authority (on the intentions of the United States Postal Service in this connection, see for example http://www.ilpf.org/work/ca/app4.htm). BT, in conjunction with VeriSign, has established the Certification Authority Trustwise (see http://www.trustwise.com).

in having the same role to play in authenticating the identity of all its citizens as "e-persons" – in the short term, at least. At present, then, individuals outside of the institutional context are left effectively "stateless" in this model. In practice, this suggests that they will continue to use the multiple approaches to access management which are currently deployed, although the adoption of trusted systems may be the most effective way of dealing with open user groups of this kind.

4.2.3 "Trusted third parties"

It is worth exploring a slightly different solution, where the user has a single identity which we assume can be certificated by a "trusted third party".

However, unlike the model above, 'rights' visas are not part of this e-person identity. To put it another way – the user is identified and authenticated by the third party, but not authorized in respect of any particular resources. Instead, the user offers her passport (her proven e-person) in order to acquire 'rights' visas. In the case of rights which are acquired by virtue of institutional membership, visas could be granted by either the licensee institution or the information provider. In the case of rights which were *not* acquired by virtue of institutional membership, visas would be granted by the information provider.

This latter model therefore offers a solution for those cases in which the user acquired rights on her own behalf, regardless of institutional membership. It assumes that everyone in the chain trusts a third party to confirm that a user is who she says she is, rather than trusting an institution to do so, as was considered in 4.2.2. Once the resource provider knows that the user is authenticated, they then turn to their own data to check that this individual has rights in respect of their resources. Whereas the model discussed in 4.2.2 placed trust for authentication and authorization in the institution, this model places trust for the former in a third party and for the latter in the provider.[30]

This has significant advantages which will be clear by now. However, we return to the issue which we referred to in the final paragraphs of the 4.2.2: how do we decide who is a trusted third party? This is a critical issue, and this model relies on its satisfactory resolution.

4.3 Identification metadata in the Passport and Visa model

4.3.1 Identification and authentication

At the most basic level, the model must enable users to offer the resource operator both an identity and a proof of that identity.

4.3.2 Authorization

That identity should have attributes associated with it. In the institutional context, users would need to express at least the following data to resource operators, and probably other data elements as well. Our intention in drafting the elements below is not to provide an exhaustive, definitive list of the data elements required to cover any institutional context. Rather, it is an attempt to scope some of the work that needs to be done if we are to standardize the way in which these data are identified.

An element is required to express the individual's affiliation to an **institution** at the broadest level. For example, it might represent a University, which is further divided into narrower sections, such as Colleges; or an LEA, in the school environment; or a commercial Group, in a

[30] And in doing so, inevitably fails to protect the privacy of the user.

commercial setting. Further, an identifier of an **institutional section or division** is required – in the examples above, to identify a College, or a school, or a branch company. **Institutional sub-division** elements must then be provided as required, in order that institutions can describe themselves at the level of granularity they require.

As far as we know, there is no standardized way of identifying institutions at these levels. However, one might presume that the university sector has developed standard identifiers in order to aid some parts of their administrative processes and it bears investigation whether these could be used in our context and mirrored more broadly.

Data elements are also required to express **cross-institutional affiliations**, from which derive authorization properties. An example in the UK HE sector is the M25 Group, a consortium of over 100 HE libraries within the region defined by the M25 motorway which was established to foster co-operation and collaboration between the member libraries in order to provide collective benefits for users.

Elements are required to identify the individual's **status** or **role** within the institution. Examples might include 'undergraduate full time', 'postgraduate part time', 'faculty part time', 'alumnus' in the University environment; 'head of department', 'teacher', 'student', in the case of schools; 'head of department', 'line manager', 'part time' in the commercial sphere. Again, as far as we know, there is no standardized way of expressing this data.

In addition, elements must express the **qualification** being pursued, if relevant (Diploma, MStud, GCSE, for example), **course** and **course options**. Although course numbers exist, we believe that they are localized rather than standardized. Further, given that the options offered by an institution change every academic year (depending on the availability of teaching staff, academic fashions, previous course performance), and bearing in mind the vast range of options offered across institutions, we are certain that, although institutions have developed their own internal identifiers for courses, they have not been standardized across institutions. This may not be a significant problem, although some sort of protocol for identification of courses will be needed. We can also imagine that users will need to express their **association with faculties**, in order to inherit the rights that belong to those faculties.

This "demographic" information is required to substitute for specific identity at the point of access.

In addition, broader, truly "demographic" information might be required, such as **date of birth**, and possibly **nationality**.

4.3.3 Financial attributes

In some cases, individuals may interact with the resource such that they receive benefits and incur costs over and above those covered by the licences to which they belong – for example, they might be enabled to print or download a number of articles. If so, the user will need to provide financial attributes which might specify, for example, that they were authorized to spend a certain amount of money from a credit account. One could imagine here that this would be an attribute of certain resources for faculty members, for instance.

5 Conclusions

It is possible to see that an implementation of the passport and visa model could circumvent some of the negative features of current practice. This would be a far from unimportant achievement, for the present situation is becoming increasingly untenable, from the point of view of users, operators and providers.

However, the model would only work if significant improvements were made – critically, since the model discussed in 4.2.2 is operating outside a technical "trusted system", it implies the placing of trust in institutions to a far greater extent than is presently the case. In turn, the model discussed in 4.2.3 requires all parties to agree that there is a "third party" whom they all trust. In addition, if passports are to be contained on smart cards, we must await a technology that is not yet common. We encourage debate as to whether the model's positive features justify the work required to implement it. At present, we are uncertain whether the model offers enough value to make its implementation worthwhile, and would welcome the views of our readers on this issue.

In order for the model to offer the realistic prospect of a single identification of users outside the institutional context which can be used to access an array of networked resources, the issues of trust must be resolved. We find it telling that none of the participants in our study have been able to find and implement a solution to this problem themselves, despite devoting considerable attention to the topic. Nor does the CNI White Paper find a solution outside the institutional framework.[31] This, then, is an issue which has attracted keen thought – but still, we believe, lacks a plausible solution.

One of the most significant conclusions of the paper must be that there is an enormous amount of work to be done in the area of user identification. Our attempt (see Section 4.3) to define elements of a controlled vocabulary to identify institutional members shows this clearly. We encourage the participation of institutions, resource operators and resource providers in the development of such a vocabulary, and we emphasize that it must applicable across all sectors and not just HE. Specifically, we recommend a more formal analysis and presentation of the metadata elements required to identify users, and stress that, if standards are to be usefully adopted, this requirement should be addressed sooner rather than later.[32]

It has been said that "liquidity substitutes for identity" in e-commerce,[33] and that the ability to pay for access to resources renders the need for identification unnecessary. This, of course, holds good for "pay per view" models for access to resources, but not for subscription models. If the problem is to be solved outside the institutional environment, it will require further careful consideration.

---

[31] See Section 5.0, *Conclusions*, in the White Paper.

[32] Some of the ground work seems likely to be done within the INDECS project in the definition of basic metadata for "interested parties" – a term which includes both individual and corporate "persons".

[33] Hughes E "A long-term perspective on electronic commerce" Release 1.0 March 1995. Available at http://www.edventure.com/release1/0395body.html

6 Acknowledgements

We would like to thank the following people for their valuable contributions to this study:

The British National Bibliography Research Fund, for providing the funding for the report
John Akeroyd (South Bank University)
David Alsmeyer (British Telecommunications)
Rosie Altoft (John Wiley & Sons)
Jos van Berkel (Swets)
Caryl Hunter-Brown (Open University)
David Brown (BIDS)
Roger Brown (Smith Kline Beecham)
Peter Burnhill (edina)
Henk Compier (Adonis)
Adrian Dale (Unilever)
Sarah Flynn (Glaxo Wellcome)
Sheila O'Flynn (Unilever)
Brian Green (Book Industry Communication)
Kevin Green (Pilkington)
Anna Grieve (Blackwells Navigator)
Mike Hannant (Royal Society of Chemistry)
Clive Hemingway (Atypon)
Helen Henderson (Information Quest)
Suzanne Wilson-Higgins (Blackwell's Information Services)
Gillian Reid-Holden (Digital Publishing Consultant)
Elspeth Hyams (Institute of Information Scientists)
David Inglis (British Library)
Mike Kelly (Open University)
Arnoud de Kemp (Springer)
Diana Leitch (Manchester University)
Andrew Lucas (Reuters)
Cliff Morgan (John Wiley & Sons)
Sally Morris (ALPSP)
John Paschoud (London School of Economics)
Jason Plent (Encyclopaedia Britannica)
Albert Prior (Swets)
Anne Ramsden (Open University)
Elspeth Scott (Glaxo Wellcome)
John Simmons (BIDS)
Jan Velterop (Academic Press)
Jenny Walker (SilverPlatter)
Robert Welham (Royal Society of Chemistry)
Jane Whitall (Smith Kline Beecham)
John Williams (TrustMarque International)
Norman Wiseman (Athens)
Robin Yeates (South Bank University)