

The Identity Project

Work Package 3: Membership

Final Report

Simon McLeish

Table of Contents

Executive Summary.....	3
1 Introduction.....	4
1.1 The Identity Project Background.....	4
1.2 Membership Life Cycles and Complications.....	5
1.3 UK and International Initiatives on Membership.....	6
1.3.1 LDAP Schemas.....	6
1.3.2 HESA and UCAS.....	7
1.3.3 Resource Licencing Initiatives.....	8
1.3.4 Federated Access Management, eduPerson, and the UK Access Management Federation	10
1.3.4 Summary.....	13
2 Credential Management and Personal Identity Management.....	14
3 Attribute Storage and Disclosure.....	16
4 Atypical Individuals.....	18
5 Prior ID Discovery and Multiple Identities.....	20
6 Virtual Organisations, Collaborative Learning, and Integration with Other Communities.....	22
7 Conclusions and Recommendations.....	23

Executive Summary

The conclusions and recommendations at the end of this document should act as an executive summary of its contents.

1 Introduction

1.1 The Identity Project Background

The Identity Project addressed the current practice and future needs of UK academic institutions in Identity Management (IdM). The IdM issues which were investigated included Grid use, Shibboleth installations of varying degree of maturity, collaborative courses and other long term inter-institutional collaborations, internal and shared dynamic virtual organisations, classes of users other than standard staff/student mix, library access schemes, and NHS involvement.

Partners in the project are:

- Cardiff University¹ (project lead partner)
- London School of Economics & Political Science² (leading the case studies work package)
- Birkbeck College³
- Goldsmiths College⁴
- Imperial College London⁵
- Queen Mary University of London⁶
- Royal Holloway College⁷
- School of Oriental & African Studies⁸
- University College London⁹
- University of London¹⁰ (associate partner, not funded by the JISC)

Each partner carried out an audit of their IdM processes, as described in the Audit process final report¹¹. A large part of this was concerned with issues surrounding membership of an institution. The project also ran a wide-ranging survey, circulated to every HE institution in the UK, which also addressed these issues; this is described in the Survey final report¹². This report is based on the information gathered through these activities.

The project started on 1 November 2006 and ended on 31 October 2007. The project was funded under the JISC e-infrastructure programme¹³.

Further information about the project generally can be found via the project web site¹⁴.

This report takes the information gained from the two major investigative work packages, the institutional audits and the survey, to obtain a picture across a range of institutions as to the current state of institutional policy and practice regarding institutional membership and IdM: what categories of users exist, and what their relationships with the institution are. It will seek to understand how these models of membership fit in with current and future legal requirements and

1<http://www.cardiff.ac.uk>

2<http://www.lse.ac.uk>

3<http://www.bbk.ac.uk>

4<http://www.goldsmiths.ac.uk>

5<http://www.imperial.ac.uk>

6<http://www.qmul.ac.uk>

7<http://www.rhul.ac.uk>

8<http://www.soas.ac.uk/>

9<http://www.ucl.ac.uk>

10<http://www.london.ac.uk>

11 <https://gabriel.lse.ac.uk/twiki/bin/view/Projects/TidpCsPmtFinalReport>

12 <https://gabriel.lse.ac.uk/twiki/bin/view/Restricted/TidpBroadSurveyReport>

13http://www.jisc.ac.uk/whatwedo/programmes/programme_einfrastructure.aspx

14<http://www.identity-project.info>

licensing restrictions. It will report upon current institution practice and will recommend actions for institutions to take to improve this aspect of IdM - improving adherence to legal and licensing restrictions - including how this compliance will impact upon IdM practice and policies.

The primary focus of this report is on institutional policies of membership - what institutions' policies define as a "member" and how members are categorised. This is an important area to investigate as institutions wishing to comply with legal requirements and licensing restrictions (amongst other things) need strict policies and definitions in place, and these could have a large impact on IdM policies and procedures. Two important areas that this includes are in understanding how policy decisions here can have funding implications for the institution at large (e.g. if a certain group are treated as institution members, they may have access to more resources and therefore licence costs may increase); and in understanding the importance of institutional buy-in at a senior level of the importance of this issue, which in turn feeds back into IdM policies and procedures.

The report is structured through an introduction exploring some of the drivers which form the background to current and future requirements for a UK HEI's management of its users, followed by sections describing credential management, attribute storage and disclosure, atypical individuals, prior ID discovery, virtual organisations, collaborative learning, and integration with other communities. The final section lists conclusions and recommendations. Unidentified quotations are taken from the institutional audit reports.

1.2 Membership Life Cycles and Complications

It may seem that membership is a straightforward business - either someone is a member of an institution, or not. The lifecycle of a user also may appear to be straightforward: a simple create, amend, and remove sequence. However, in the higher education sector, this simple sequence will occur only in a minority of cases. For example, an undergraduate may have some rights to access systems even during the application process (e.g. a site which permits tracking of the progress of their application), may have access rights which depend on which courses they take (e.g. to a VLE), may have a gap in studying, and may have access rights as an alumnus (e.g. alumni email accounts). After apparently leaving the institution, they may return as research students and eventually staff. As staff, they may have differing levels of access to different internal and external systems at different times (e.g. the finance system at times when they manage a budget code). The boundary between staff and student is not precise; in many institutions, staff can take courses, and research students are often on the payroll (such as situation is described by the individual taking on multiple roles).

Not only is the boundary between staff and student blurred, but the boundary between member and non-member is too. A key example is that of the library walk-in user, who has rights to access material (often both physical library resources and external electronic resources) through physical presence in the building - and the requirements about how a member of the public or even a student at another institution becomes a walk-in user differs between institutions (see the entries in the M25 consortium Visit a Library service¹⁵ for examples). Other classes of user with ambiguous status include alumni and associates (generally an ill-defined role).

¹⁵http://www.inform25.ac.uk/AET/cgi-bin/aetaccess.pl?display_search_form=yes&show_first_only=yes

1.3 UK and International Initiatives on Membership

This section describes a variety of different approaches which have informed IdM and membership discussions in particular in UK HE institutions. This selection is not intended to be comprehensive, but to provide a background to the problems and potential solutions regarding membership and user attributes in particular.

1.3.1 LDAP Schemas

The discussion so far has been single institution centred. However, membership has always had some outward facing aspects (resource licensing; rights at other institutions, etc.), and these are currently becoming increasingly important. The reason for this is basically the advent of Federated Access Management (FAM), which generally uses information provided by an institution about its members to determine their rights of access to external resources. Such information is considered to be a collection of attributes of the user, and is usually stored in databases known as directories maintained by the institution.

Many directories use LDAP (Lightweight Directory Access Protocol). This is based on the earlier X.500 DAP, which required Open Systems Interconnection, a networking protocol eventually superceded by the TCP/IP protocol used by the Internet, and was first developed in 1992 by the University of Michigan to provide a mechanism for accessing X.500 DAP directories over TCP/IP. The current version of LDAP (version 3) is defined in RFC 4511¹⁶. Less technical introductions to LDAP can be found on the Identity Project links list¹⁷.

Apart from defining the principal method used to transport attribute information, LDAP also uses a set of standards for the definition of attribute information, known as "schemas" and "object classes" (object classes are collections of attributes; schemas can contain both attributes and object classes). Schemas do not have to be about people: there are ones which describe organisations, some which describe computing equipment, and so on. Some of the earliest international initiatives on membership and user description still in common usage are important LDAP schemas. Ones which are frequently pre-installed in LDAP directories include the following.

- **country** and **locality**: describe places
- **organization** and **organizationalUnit**: describe institutions (in this context) and departments
- **person**, **organizationalPerson** and **inetOrgPerson**: describe an individual and how they relate to an organisation; use attributes from the two preceding lists

It should be mentioned that these attributes are also commonly used in X.509 certificates used for authentication.

More details on these and other schemas can be found at <http://www.oav.net/mirrors/LDAP-ObjectClasses.html>.

LDAP attributes are defined, as far as usage and semantics are concerned, by English language descriptions; for example, the `displayName` attribute from `inetOrgPerson`¹⁸ is defined as:

When displaying an entry, especially within a one-line summary list, it is useful to be able to identify a name to be used. Since other attribute types such as 'cn' are

¹⁶<http://tools.ietf.org/html/rfc4510>

¹⁷ <https://gabriel.lse.ac.uk/twiki/bin/view/Projects/IdentityDocs>

¹⁸<http://www.faqs.org/rfcs/rfc2798.html>

multivalued, an additional attribute type is needed. Display name is defined for this purpose.

Such definitions are useful guides, but do not absolutely guarantee similarity of interpretation between different directories.

LDAP directories are generally used for two purposes: as an email address book service (many LDAP servers are part of an email system), and to provide data for applications such as Shibboleth which pass attributes to external systems. The database used by the directory is usually fed from other systems which are less accessible from outside, such as MIS databases, and the information included in an LDAP directory is dependent on what is found in its source databases. This may well have been re-purposed from user information gathered for other purposes, and without care in doing this, it is possible that the data itself does not precisely fit the definition of the attribute used to expose it.

1.3.2 HESA and UCAS

In the UK, the Higher Education Statistics Agency¹⁹ (HESA) has a responsibility for collecting information from HE institutions for the production of statistical information about the sector since 1993. For this purpose, it is important that definitions agree from institution to institution, as this information is used by individual institutions and by central government for planning and policy.

HESA have published information describing those members of an institution in whom they have an interest. The information is very detailed, and a lot of it is not likely to be relevant for purposes of IdM (e.g. the field LANGPCNT, "used to indicate the percentage of the module that is taught through the medium of a Celtic language"), or would constitute a severe privacy breach if revealed (e.g. ethnicity). However, returning data to HESA is a statutory obligation for UK HE institutions, which means that there must be appropriate mechanisms in place for gathering such data and ensuring its reliability. Institutions must, in particular, have made decisions about how to discover relevant information and have procedures for creating reports for HESA using the mandated data structure.

HESA does not define what precisely is meant by a student, though the data structure indicates that a student must be taking at least one course (though that begs the question: what constitutes an acceptable course when there are short courses, foundation courses and diplomas?). There is a more formal definition of staff, which is:

All staff for whom the institution is liable to pay Class 1 National Insurance contributions and/or who have a contract of employment with the institution must be included in the record.

This may exclude some honorary and visiting staff who may be awarded staff privileges for certain purposes, such as access to library resources.

UCAS²⁰ provides application services across a range of subject areas and modes of study for UK universities and colleges. This means that UCAS is able to provide a large dataset of information about applicants, as can be viewed at http://www.ucas.com/about_us/stat_services/stats_online/data_tables/. This data is available to an institution when a student takes up a course. Many of the data fields here are now aligned with

¹⁹<http://www.hesa.ac.uk>

²⁰<http://www.ucas.com/>

those used by HESA.

1.3.3 Resource Licencing Initiatives

Since access to third party resources is not just a major use of user information but also one which exposes it to the outside world, it is important for institutions that their use of categories such as student and staff matches that of the licenses which permit access to these resources. In this section, we look at some of the important models for agreements in UK HE and see how these define authorised users.

Licence	Authorised users	Definition of Walk-in Users
JISC Model Licence ²¹	<p>Current members of the staff of the Licensee (whether on a permanent, temporary, contract or visiting basis) [,/and] individuals who are currently studying at the Licensee's institution, who are permitted to access the Secure Network from within the premises of the Licensee and from such other places where Authorised Users work or study, including without limitation halls of residence and lodgings and homes of Authorised Users, and who have been issued by the Licensee with a password or other authentication.</p> <p>Some licences may be for a more limited group, such as students enrolled on a specific course.</p>	<p><i>Walk in Users Institutional Premises</i> Persons who are not a current student, member of staff or a contractor of the Institution, but who are permitted to access the Institution's information services from computer terminals or otherwise within the physical premises of the Institution ["Walk-In Users"] are also deemed to be Authorised Users, only for the time they are within the physical premises of the Institution. Walk-In Users may not be given means to access the Licensed Work when they are not within the physical premises of the Institution. For the avoidance of doubt, Walk-In Users may not be given access to the Licensed Work by any wireless network provided by the Institution unless such network is a Secure Network.</p> <p><i>Walk In Users - Library Premises</i> Persons who are not a current student, faculty member or an employee of the Sub-Licensee, but who are permitted to access the Sub-Licensee's information services from computer terminals within the Sub-Licensee's Library Premises ["Walk-In Users"] are also deemed to be Authorised Users, only for the time they are within the Library Premises. Walk-In Users may not be given means to access the Licensed Materials when they are not within the Library Premises.</p>
Chest ²²	Currently registered students,	<i>Walk-in User</i>

²¹<http://www.ukoln.ac.uk/services/elib/papers/pa/licence/Pajisc21.html>

²²<http://www.eduserv.org.uk/chest>

	<p>faculty members or employees of the licensed institution who are authorised by the licensee to access the licensee's information services whether from a computer or terminal on the licensee's secure network or off site via a secure access management system.</p>	<p>A person who is not a currently registered student, faculty member or employee of the licensed institution but is permitted by the institution to access the secure network via a computer or terminal within the Library premises is deemed to be an authorised user but only for the duration they are within the Library premises. Institutions that provide access to networks, and users who benefit from that access, should regard it as normal to require an individual identity.</p> <p>Walk-in Users may not be given means to access the licensed work when they are not within the Library premises.</p> <p><i>Unidentified Visitor</i> This term refers to a general public user who is an unknown, anonymous user not a currently registered student, faculty member or employee of the licensed institution and who does not have remote access to resources.</p>
NESLi2 ²³	<p><i>Authorised Users</i> means individuals who are authorised by the Licensee to access the Licensee's information services whether on-site or off-site via Secure Authentication and who are affiliated to the Licensee as a current student (including but not limited to undergraduates and postgraduates), member of staff (whether on a permanent or temporary basis including retired members of staff and any teacher who teaches Authorised Users in the United Kingdom) or contractor of the Licensee.</p>	<p>As the JISC Model Licence above</p>
Liblicense Standard Licensing Agreement ²⁴	<p><i>Persons Affiliated with Licensee.</i> Full and part time students and employees (including faculty, staff, affiliated researchers and independent contractors) of Licensee and the institution of</p>	<p><i>Walk-ins.</i> Patrons not affiliated with Licensee who are physically present at Licensee's site(s) ("walk-ins").</p>

²³<http://www.nesli2.ac.uk>

²⁴<http://liblicense.ukoln.ac.uk/>

which it is a part, regardless of the physical location of such persons.
--

It can be seen that these differ slightly in their definitions of both authorised users and walk-in users (who may be authorised access to certain resources). It is clearly a requirement that information about the individuals affiliated to an institution should make it possible to comply with the clauses of these and other licenses, and where the definitions differ (e.g. the NESLi2 licence including retired staff excluded in the JISC model licence) the institution will need to restrict access depending on the licence. The situation can become particularly complex where multiple access routes exist to the same resource (e.g. journals accessible through more than one content aggregator) and where these have differing licence agreements.

The HAERVI Project²⁵ has produced a report²⁶ in September 2007 with recommendations on best practice in dealing with walk-in users. Many of its recommendations address issues already mentioned, including:

- JISC collections and Eduserv CHEST should clarify whether the term walk-in applies to location of the visitor within library or institutional premises. the latter would be the preferred form which would bring most flexibility to visiting users, e.g. visiting academics with a temporary desk in a host department.
- Institutions should consider HE visitors requesting access to their electronic resources within the context of their overall policy for facilities granted to visitors.
- Institutions should analyse the different ways in which visitors are granted access to your network and e-resources, and rationalise these routes, if necessary.

1.3.4 Federated Access Management, eduPerson, and the UK Access Management Federation

A growing use of attributes, and one which exposes an institution's information about their users to the external world, is for federated access management (FAM). This entails access to material at a service provider with authentication provided through an "identity provider" hosted by an institution (as opposed to previous models, where authentication was centralised (classic Athens) or provided by the service provider itself. The federated access model is generally (though not necessarily) role based, with the identity provider giving information through attributes about roles applicable to the user (e.g. the information that they are an undergraduate student studying a particular collection of courses) and the service provider using this information to restrict access appropriately. It needs to be mentioned that this mechanism is not currently used with great sophistication in general, with the role information of membership being enough to gain access to the vast majority of resources. However, it is likely that more roles will be used as time progresses, so that (for example) students have access to specific resources applicable to their course selection. This makes the accuracy of the attribute information of primary importance, particularly regarding the removal of obsolete information.

In the UK HE community, the most commonly used FAM software is Shibboleth²⁷. It uses bilateral trust agreements between identity providers and service providers to allow the processes of authentication and authorisation to be carried out by the respective parties, and it is usual for these agreements to be aggregated to constitute membership of a federation. Shibboleth is agnostic

²⁵<http://www.ucisa.ac.uk/haervi/haervi.aspx>

²⁶<http://www.ucisa.ac.uk/haervi/agree/>

²⁷<http://shibboleth.internet2.edu>

about the attributes it passes to service providers, relying on bilateral agreements and federations to define which attributes are passed and what their values mean. Shibboleth also allows anonymised role based access, where the roles are all that is passed about a user and (unless the attributes defining the roles include some which match to a single individual, such as a personal email address) make it impossible to directly identify the individual who is requesting access to a resource protected by the service provider. While FAM is commonly considered in relation to access to third party electronic resources, it also is likely to play an increasing role in cross-institutional collaboration, both shared courses and distributed research groups. This was, in fact, the original use case that drove the development of Shibboleth.

The attribute schemas used in Shibboleth are usually reused LDAP object classes. One new LDAP object class particularly associated with Shibboleth is eduPerson²⁸. This references (and slightly redefines in some cases) attributes from inetOrgPerson, orgPerson, and person, and introduces a range of new attributes relevant to role based access, particularly where this is anonymous. Since these will be new to many institutions when they install Shibboleth, we give some notes on these attributes here.

Attributes	Notes
eduPersonAffiliation, eduPersonScopedAffiliation	Defines the nature of the relationship between an individual and an institution. It has a short list of permitted values: faculty, student, staff, alum, member, affiliate, employee. (Walk-in user is currently being considered for addition to this list.) There are many classes of users which do not fall neatly into this list from UK HE (see Section 2.1, User Categories), others in frequent usage for access decisions which subdivide the permitted values (e.g. students into undergraduate and postgraduate), and others which do not mean the same thing in the UK HE context as in the US context for which the list was originally proposed. The scoped affiliation includes information about the context of this information (generally, but not necessarily, the institution).
eduPersonEntitlement	An assertion that an individual is entitled to access to a particular resource (or collection of resources, such as those protected by the service providers of a specific Shibboleth federation). While this attribute goes against the ideas behind role based authorisation (because it means that the authorisation decision is made by the identity provider rather than the service provider), it can prove useful in some contexts.
eduPersonNickname	Self evident, but little used and not likely to be relevant for IdM.
eduPersonOrgDN, eduPersonOrgUnitDN, eduPersonPrimaryOrgUnitDN	The identifier for the organisation, or (principal) organisational unit in the institution's directory. Not likely to be relevant for Projects.IdM.
eduPersonPrimaryAffiliation	Specifies the main relationship between an individual and an institution (choosing between the possible multiple values for eduPersonAffiliation). Currently, where multiple roles apply it is usual for an individual to get the highest level of access to which they are entitled, though this may not always be appropriate (e.g.

²⁸<http://www.educause.edu/eduperson/>

	when a member of staff is also a student, they should not have staff level access to VLE material on courses they are studying as a student). However, it is not clear how this attribute would resolve such issues, as the individual concerned will need to be able to access material with affiliations other than their primary affiliation.
eduPersonPrincipalName	An identifier assigned to the individual for the purposes of FAM. Reveals the user's identity.
eduPersonTargetedID	As above, but anonymised: this allows the service provider to identify a user as one who has previously used the service, without actually revealing their identity.

The eduPerson object class is managed by MACE-DIR²⁹, the directories working group of Internet2. This group also works on the related eduCourse³⁰, a draft schema (published in 2005) and recommendations for good practise in the identification of objects related to courses. This is more abstract than the data structure used by HESA for students, but covers much the same ground. It has also in the past worked on international equivalents and variations on eduPerson, producing A Comparative Analysis of Collaborative Public LDAP Person Object Classes in Higher-Education³¹ in 2005. The coverage of this report is similar to Sections 1.3.1 and 1.3.4 of this report, but is much wider in scope.

For UK HE institutions, the requirements made by the UK Access Management Federation³² will be the main focus of IdM work aimed at enabling FAM. The federation is officially agnostic about technology, though the reference implementation of Shibboleth from Internet2 is currently the software used by the vast majority of federation members (just under 95% on 6 October 2007). Like Shibboleth itself, the federation does not make any specific requirements about attribute usage, though it suggests³³ that core attributes from the eduPerson object class be used as these are likely to be required of identity providers by service providers to gain access to resources. There are also recommendations about de-provisioning expired users and accuracy of data in directories.

Neither Shibboleth, eduPerson, or the UK Access Management Federation give fully comprehensive advice on how institutions should decide which users should have particular values of attributes associated with them. This is in large part because institutions are not all the same, and have slightly differing ideas about (for example) when someone is a member. Currently, the major way that this issue arises is because institutions want to be able to use attributes correctly but cannot find appropriate guidance; we suspect that publishers are not generally concerned provided that the number of "members" granted access by an institution approximately matches the number paid for by the institution.

As Shibboleth federations proliferate, an issue which is beginning to be discussed is about inter-federation co-operation, especially the possibility of permitting identity providers from one federation to access service providers in another. The first International federation peering workshop took place in September 2007 in Prague (see the JISC FAM blog³⁴ for more information). The major issues are procedural and legal rather than technical (e.g. ensuring that conditions of membership are not broken through access by members of other federations, or that differing privacy laws are appropriately applied across borders), and work on these is likely to provide

²⁹<http://middleware.internet2.edu/dir/>

³⁰<http://middleware.internet2.edu/courseid/docs/internet2-mace-dir-courseID-eduCourse-200507.html>

³¹<http://middleware.internet2.edu/dir/docs/draft-internet2-mace-dir-higher-ed-person-analysis-latest.htm>

³²<http://www.ukfederation.org.uk/>

³³<http://www.ukfederation.org.uk/library/uploads/Documents/technical-recommendations-for-participants.pdf>

³⁴<http://involve.jisc.ac.uk/wpnu/jam/2007/10/01/potential-for-access-resources-across-international-borders/>

further drivers towards standardisation of membership and user metadata on an international scale, which will clearly have an as yet unpredictable effect on UK HE IdM. A JISC study is under way on the feasibility of a cross-jurisdiction Common Access Management Federation Agreement³⁵ with a blog at <http://jisc-legal-fact.blogspot.com/> from the beginning of October.

1.3.4 Summary

It is clear from these various initiatives that UK and international initiatives relevant to membership are developing rapidly, and moving towards greater standardisation. However, there are still contradictory requirements in many areas, and it is incumbent on any institution to pay careful attention to the way that user metadata is derived and used. This is particularly the case when an institution becomes a participant in FAM.

³⁵http://www.jisc.ac.uk/whatwedo/programmes/programme_am_transition/fedpolicy.aspx

2 Credential Management and Personal Identity Management

There is considerable variety in the mechanisms by which users are provided with credentials for access to the various systems to which they have rights.

Credentials for the typical staff and students of an HEI are created automatically by every Identity Project partner, based on data entered into a Human Resources or Registry System. Frequently, such an account is activated by the users themselves, by the sending of an email to a personal email address requiring them to access a web page and enter a combination of personal information unlikely to be known in combination to a third party (e.g. date of birth, a unique identifier for their application, etc.). This type of process is more likely to happen for students than for staff; Human Resources departments tend to enter details on behalf of successful applicants.

Where accounts, credentials, and user metadata for different classes of user are the responsibilities of different departments, there is an issue of ownership: there is no overall control and it becomes possible for differing sets of requirements to be applied to the different classes rather than a coherent central policy. Many different software systems throughout an institution will have the ability to create and use their own credentials for users, and it has in the past generally been easier to use these and get the users to remember multiple sets of credentials than it is to persuade software (which is often effectively a "black box" to the institution to some degree at least) to use credentials generated and verified elsewhere.

Most of the Identity Project partners provide each user with some kind of unique identifier which is intended to remain the same even if other attributes which are specific to the individual change, through the duration of their entire relationship with the institution. (The obvious example here is change of name, as when this happens email addresses and even network userIDs may be changed to reflect this.) Identifiers may also be associated with individuals outside the institution, who would not be considered to be members: those with a financial relationship to the institution, for example. However, in many places, different systems which contain information about different kinds of users often use different identifiers, though in others a useful property of a universal identifier is that it can be used to link records in diverse sources. The reasoning behind this is partly to help with prior identity discovery (as discussed in section 5) and partly to ameliorate the ownership issues just alluded to.

A major issue with the management of users of all kinds is de-provisioning (expiry). Clearly, having users remain in a system and retain rights that they should no longer possess is relying on the honesty and apathy of the former user to avoid breaches in licensing and other similar requirements. There are two options: removing the user entirely or changing their rights. Generally, institutions wish to retain records of former users for many reasons, so it is likely that the user will remain in some databases, though they may be removed from others (e.g. kept in the MIS system but removed from LDAP directories used by Shibboleth). There is the additional question of re-use: in some cases, it may make sense to re-assign some parts of the components which make up the user's identity within an institution to other users (e.g. access to role based email accounts such as *estates.management@inst.ac.uk*), and in others they are re-assigned because that is what is done (e.g. personal email addresses - which is usually justified to avoid users with common names having complex email addresses, though this is less of a problem for them now than it used to be as they will usually have accounts on web sites such as Hotmail or Facebook and will rarely be the first to choose johnsmith as their userID). In every system which lists the user, some action needs to be taken, and this proves problematic for the majority of institutions.

- There may be different re-use policies for different systems (e.g. network logins are never re-used, whereas email addresses often are), and some record will need to be kept for a long time for both students and staff
- Atypical users often have anomalous credentials (e.g. manually created network accounts separate from automated batch processes) which do not automatically expire and may therefore never be expired. A good practice recommendation is that any account should have an expiry date set on creation, though some thought needs to be given to this (e.g. ensuring that default expiry dates such as six months from creation fit the needs of the user needing the account)
- Students and staff will not all follow the expected path (e.g. students may take exam resits, staff may have contract extensions or leave early) and information about this needs to be passed to all systems which would deprovision them

Where details about an individual connected to the institution need to be changed (e.g. new address details) it can be time consuming if these changes need to be made by proxy (e.g. by submitting information to the personnel department for them to update a database). On the other hand, allowing users to make changes themselves runs the risk of inaccurate data being supplied, which can prove problematic (e.g. where students are allowed to choose a preferred, non-institutional, email address and forget to update this when it changes). Some of the Identity Project partners permit users to update at least some of the details held by the institution, typically through a web portal, and others are considering doing so. As staff and students are typically listed in separate databases with different front end software, it is possible that they will be able to make different types of changes; at some institutions, students have greater abilities in this area than staff.

3 Attribute Storage and Disclosure

In this section, we consider attribute stores; that is, databases which contain significant data about individuals made available to other systems. Within an HEI, some stores are considered to be authoritative sources of information about users (e.g. Human Resources and Registry systems), while others (such as LDAP directories) act as conduits to access such information (often in a read-only fashion, though occasionally allowing modification of details passed back to the source) without giving direct access to the authoritative source, and frequently combine data from several sources. There are also many stores which fall between these extremes, which add their own information to data obtained from elsewhere (e.g. Active Directory, which adds information about network usage to user metadata, or the Library System, which often handles user data for walk-in users separately from members of the institution). The relationships between these, and the large numbers of smaller systems which contain data about users are complicated, and it is important to manage these carefully in order to reduce duplication of effort and potential issues particularly with de-provisioning not being universally propagated.

One issue with keeping the authoritative sources of information securely hidden away that was noted is that "those responsible for the creation of data are not necessarily those with an interest in the use of particular attributes" - those working on source data may not even know the purposes for which other directories might eventually use the data. This is one aspect of a general problem with data re-use, which arises because data is created or managed for one purpose and used for another, and the two may not be entirely compatible (e.g. if the definition used to define a group of users is different from that expected by the re-user).

Directories are generally fed data from a variety of sources, which include:

- custom built or purchased systems for managing users (including ones where users have access to maintain their own data)
- proprietary MIS databases
- personnel and payroll systems
- financial systems (particularly those which permit students to purchase goods and services on campus)

Where there is no universal unique identifier, some other means has to be found to link records describing the same individual from different sources.

Major update methods for directories include:

- flat files provided to/from other systems on a periodic basis as exports/imports (traffic may be one way only, or may flow in both directions)
- event triggered automatic updating scripts for individual changes
- manual updates/downloads, either regular or triggered by alerts from the sources

Security issues were noted in some instances where sensitive information, including passwords, was passed in clear text when data was transferred from sources to directories.

There is a tendency for academic and ancillary departments to make local copies of user records (or their own records duplicating data held in central sources) for particular needs. Examples cited include:

- academic departments holding details of temporary staff who might be invited back, for contact details and to maintain a record of payments - also separately from the central finance system)

- information relating to student disciplinary action

A principal use for information obtained from directories by applications is authorisation, both internal usage (e.g. to populate permission files for intranet facilities) and passed to third party service providers as part of FAM. The other major use for directories is to provide address searching for email contact information (which is the reason why they tend to be packaged with email server software). Other directories in the sense defined above include many library systems and alumni contact databases.

Several Identity Project partners indicated that there were departments in their institutions who maintained independent networks and user management processes. This was usually for historical or technical reasons (e.g. preferring UNIX platforms rather than Windows). Integration with central sources of data tended to be mixed where this was the case. With the introduction of FAM, this could escalate into a serious problem.

Disclosure of information about users is subject to legal controls such as contractual agreements and the Data Protection Act. This often makes institutions cautious about revealing data, particularly of staff members. However, measures to prevent erroneous disclosure of information are often ad hoc (other than in particularly sensitive cases such as police requests) and this will mean that important requirements (such as whether the individual needs to be informed, or give permission) may not even be considered.

Users are generally not greatly concerned or knowledgeable about the ways in which data held by the institution about them is used (in particular, they tend to have a perception that data is held about them on one system within the institution). As one partner indicated in their audit report, in general, users typically remain sanguine and/or ignorant as to if and how attribute information is transferred. Most users seem to trust their institutions to make sensible use of personal data at the moment, with concerns being raised by IT staff before general users become aware of them (e.g. during the planning stage of a portal at one partner to allow users to edit data contained in institutional directories, issues deriving from making student photographs held by the institution available on the web were discussed).

However, there is also much "under the surface" transfer of personal information in addition to the large scale data flows discussed above, which may be small scale but could compromise security (as has happened in other sectors: see for example a news story³⁶ from October 2007). Such transfers of information may be entirely unofficial or part of established business processes. Data is transferred via email (can you give me Smith's mobile number?), via paper forms, vocally (e.g. when someone physically visits a Human Resources Department and asks a member of staff there for payroll details) in shared and personal databases (e.g. contact details for interviewees for a research project), and the individuals involved need to have some level of understanding of their responsibilities regarding security and privacy (e.g. for the HR department visitor, it is important to consistently require as much information about identity as would be needed for transfer of this information electronically). It should be noted that there are still instances where paper forms are used to supplement electronic data storage, particularly for finance systems (where countersignatures may be required from a senior member of staff before access is granted) and library systems (where signatures may be required to indicate the user's acceptance of copyright restrictions). Several partners emphasised the importance in this context of educating staff in their responsibilities, and trusting them to carry them out in a professional manner.

³⁶http://www.computerworld.com/action/article.do?command=viewArticleBasic&taxonomyName=storage&articleId=9042001&taxonomyId=19&intsrc=kc_top

4 Atypical Individuals

While it is generally speaking possible to lump together standard staff and students, there are likely to be sub-divisions of these groups which affect resource access. The most important of these are vertical divisions into courses (as managers, teachers, markers, administrators, current and past students). Also well understood are divisions into academic and non-academic staff, and postgraduate and undergraduate students (with the former also subdivided into taught and research). Slightly less obvious divisions would include those based on accommodation (students who are residents in a particular hall of residence will have better access rights than other students) and on progress through a degree programme (e.g. students taking a year abroad for language degrees, or having extra rights as final year students - though the latter is now uncommon).

There are many other groups of users at many institutions, with differing rights (though more than a member of the public). The number of such groups tends to increase with size and complexity of the HEI, and they are usually less well defined than staff and student groups. Common categories include:

- departed staff (emeritus professors and retired staff generally) and students (alumni); some departments in Identity Project partners are looking at offering professional development courses to alumni, blurring the boundary between current and former students
- honorary academics
- prospective students
- external examiners and members of governing bodies
- conference attendees
- short course and summer school attendees
- contractors, temps, and employees of third party suppliers (e.g. building maintenance staff, catering staff, and cleaners)
- students' union staff and sabbatical officers
- staff for research centres
- external members of facilities connected with the HEI (e.g. sports centre members), especially the library (see below), and members of organisations involved in community projects associated with the HEI
- visitors who require access to buildings beyond that permitted to normal members of the public (e.g. open day visitors, exhibition visitors, graduation ceremony guests, guests at functions hosted by the HEI or by hirers of HEI facilities, the emergency services)

NHS use adds many more extra categories of users, and the specific implications of this are discussed in the project's NHS report³⁷. Similarly, there can be many different groups of library users, usually with similar rights but slightly different requirements for proof of identity, from different agreements with other libraries and institutions. For institutions in the area surrounding London, details of such agreements can be found via the M25 consortium Visit a Library service mentioned earlier.

Contractors and temps provide several examples of anomalous and potentially insecure practice, and yet are often treated in an ad hoc manner. Issues can include re-use of credentials, whether deliberate policy or a pragmatic work-around (e.g. network usernames and passwords passed from one temp to the next), the potential for access to material they shouldn't be able to access (e.g. if a student from the institution takes part time work as a cleaner via a contractor), and disregard for

³⁷ <https://gabriel.lse.ac.uk/twiki/bin/view/Restricted/TidpNhsFinalReport>

standard procedures (e.g. leaving normally locked doors propped open to give easier access). Technical contractors often require high level access to secure systems (e.g. administrator level access to servers) which must be carefully controlled. Additionally, the institution will need to rely on the contractor's management to ensure that IdM is carried out; for example, when an employee of a consultancy firm used by the institution leaves that employment, the institution will need to review where they had access even if the relationship with the firm continues, and this is reliant on receiving accurate information from the firm.

However, all the atypical user categories can cause problems, because they do not fit into the established procedures for maintenance of user accounts and metadata. (As one Identity Provider partner report put it, there is no standard procedure for dealing with non-standard users .) Examples of this include retired staff, where the institution decides to exercise the right given in some licences for them to access electronic resources, but where they do not have standard institutional credentials. De-provisioning can cause particular problems: where a person's relationship with an institution is not well defined in the first place, the cessation of the relationship is likely to be hard to trace.

This means that the ad hoc methods used to provision and de-provision atypical individuals can lead to problems with accountability (there is a lack of documentation, and the process may be carried out in a way which is opaque to central services, without a proper audit trail, which makes it hard to trace such an individual in cases where rights are abused, and may even make it impossible to know how many people currently hold certain access rights) and security (the users may be set up in an inappropriate manner, if the person who does it is unfamiliar with the institution's standard procedures, and they may not be de-provisioned when they need to be, which means they continue to be able to exercise rights they no longer have).

It is also possible for standard types of users to be problematic, when they require non-standard provisioning. An example of this is a partner with a student card that is occasionally issued to staff members, who will not be on the database which is usually used to provision the card issuing system. (In this case, card are issued with anomalous numbers to indicate the non-standard nature of the issuance.) Students dropping out from their courses are another cited example. Issues also arise when standard procedures are ignored, e.g. when academic departments hire research assistants without going through human resources.

5 Prior ID Discovery and Multiple Identities

Prior ID discovery is the process of checking whether an individual has a previous history of contact with an institution. This process is time consuming and/or not fully accurate (depending on how automated it is), but there are several reasons to carry it out:

- It may be considered useful to use only a single set of credentials for an individual, over at least a limited time period or just to avoid duplicates of currently active accounts (and this is often particularly important where students study part time and take time off in between courses)
- It helps detect problems where multiple routes have been used to create identities (e.g. academic departments creating staff details independently from central systems)
- It may be considered useful to track the history of an individual's relationship with an institution, for marketing purposes, or to analyse student/job applications
- Data may come from several sources (particularly prospective student contact information) and therefore be very likely to contain duplicates
- Where there are multiple records, changes to personal information may not be made to the correct copy. (On the other hand, synchronising data from multiple relationships may cause other problems, e.g. when a former student changes their name, updating all the information about them may lose the connection between the HEI's record of them as a student and the name on their degree certificate.)
- A desire that individuals' identifiers really are unique
- It may provide useful marketing information (e.g. that summer courses are good opportunities to promote the institution)

Where a previous relationship with an individual is found, it is less expensive and time consuming to re-use existing details from the start than it is to remove duplicates if discovered at a later date. Where persistent unique identifiers are used, the problems caused when these turn out not to be unique seem to be considered more severe than where there are multiple identities but no persistent identifiers. Effectively, though, the problems are the same: rights and roles are easily given to the wrong account, causing problems with access to other systems, and the user involved needs to ensure that they are identified with the right one of their multiple identities for any action they wish to carry out. Some of these problems may prove particularly intractable; a cited example is a VLE system which uses the persistent identifier for an individual as part of the unique identifier for the user in its local database, but which is then impossible to remove when the identifier proves to be a second one for the individual concerned. It is clearly important that systems are able to merge identities which are discovered to be duplicated.

Most of the Identity Project partners carry out at least some prior ID discovery. Restrictions and problems include:

- new users are only checked against existing accounts in central systems
- some systems are not checked (e.g. prospective student contact listings are allowed to contain duplicates, which are winnowed out when an application is made)
- some types of contact may never make it onto central systems (e.g. registered library walk in users) and subsequent contacts therefore are unrecognised
- there are no general procedures to carry out prior ID discovery, so that where it does happen it is ad hoc (e.g. depending on the user noticing a problem leading to a retrospective fix)
- there are no general procedures to carry out prior ID discovery, leading to some individuals being checked multiple times at different points in the credential allocation procedures, duplicating effort

For automated discovery, multiple fields are generally used. The standard set appears to be:

- Date of birth
- Name (either full name or surname and initial)

It is generally the case that a false positive (identification of a link where none exists) is more serious than a false negative (missing an existing link), and systems tend to be configured with that in mind.

Non-automated discovery tends to happen only when an individual makes a prior relationship known (e.g. a successful job applicant makes it known to HR that they are an alumnus, or where multiple identities cause problems).

Where a user has multiple roles within an institution (e.g. both staff and student), it has generally been the case in the past that they have been issued with multiple sets of credentials with different rights. This is still true for almost all institutional staff with an identity associated with the NHS (see the NHS report previously mentioned). However, in most cases, institutions are moving towards amalgamation of role information so that multiple roles are associated with a single identity. Where an individual's roles change (rather than needing to be duplicated, as is the case when a student becomes staff), delays in activation of their new role can lead to a complete loss of access. There are also complications with attribute information (e.g. systems may expect a user to have only a single departmental affiliation, which is not always the case), especially where such information is embedded into credentials (e.g. a department identifier as part of the network username).

Alumni databases often contain multiple records for individuals, because there are usually multiple ways in which records can have entered these databases in the first place (automatically from registry when students leave - which may happen multiple times when students study more than once at the same institution; from records held by departments; when alumni contact the institution; from information passed on by other alumni; etc.). The problems are mainly historic, from before current systems of recording data about users at an institution were put in place (particularly unique identifiers), but it is still possible for a new duplicate record to be created, as is evidenced by Identity Project partners who carry out regular checks through their alumni database records for duplicates.

6 Virtual Organisations, Collaborative Learning, and Integration with Other Communities

Currently, identity management for these areas is handled by whatever means exists within the technology used by the virtual organisation or learning consortium (e.g. a VLE, which may permit external users to be created). In the absence of much evidence from the Identity Project audits, it is suggested that much of this identity management in this context is not regarded as such by those who carry it out (e.g. with collaborating authors in a research group emailing each other with draft papers, because they know who the other authors are). Shibboleth was mentioned as a key technology for future work in with virtual organisations and in collaborative learning, both within UK HE and extending into other communities.

Some HEIs also provide institutional accounts for external collaborators in research projects when requested by research groups to do so. In this instance, the group would act as a sponsor, and is responsible for appropriate use of the institution's facilities.

One specific problem that was identified in this area was with students from partner institutions abroad who are taking courses at a UK HEI. This is that paper based requirements (e.g. signed forms from the students) will take a lot longer to arrive in the UK than electronic information. Where the delay is particularly lengthy, this may lead to accounts being created and becoming accessible to the students before the paperwork arrives, and the security of the process is entirely reliant on the trust between the UK HEI and their overseas colleagues until it does.

This type of collaboration is growing in terms of scope (e.g. more international teaching and research collaboration), and is making increasing use of official central services and resources (e.g. VLEs).

Collaboration via the National Grid Service is covered in a separate report³⁸. The NHS has already been mentioned several times, and is covered in a separate report.

Other services which have been listed in this context in Identity Project partner reports include the Janet Roaming service³⁹.

Managing identities across institutional boundaries is described as an area where Identity Project partners have little experience, but is identified as an area of growing need.

38 <https://gabriel.lse.ac.uk/twiki/bin/view/Restricted/TidpNgsFinalReport>

39 <http://www.ja.net/roaming/>

7 Conclusions and Recommendations

In this section, we list some principles which can be drawn from the preceding discussion, which are worth bearing in mind when planning Identity Management activities.

- *User Categories* - HEI membership is much more complex than a naive analysis would suggest, and there many more categories than envisaged by (say) the eduPerson schema, with different rights. Not only that, but even apparently well understood terms such as student and staff are more fuzzy than might be expected. The precise definitions of these differ between HEIs, but more homogeneity of definition is likely to be required by licensing in the future.
- *Credential Management* - Automated processes used for common categories of users appear to work well and securely. Some departments may, however, act independently from the central administration processes, which can lead to problems with integration of data about department members with the rest of the institution.
- *Attribute stores* - HEIs contain large numbers of attribute stores of various kinds which are used for different purposes. These have complex interrelationships which need to be managed carefully to avoid duplication of effort and insecure methods for transferring data.
- *Unique Identifiers* - Many institutions found that a universal unique identifier for individuals worked well for synchronising data from different sources and for resolving issues with users who have multiple relationships with an institution.
- *User understanding* - Users are generally not greatly concerned or knowledgeable about the ways in which data held by the institution about them is used. This seems likely to change if there are scandals about data exposure in UK HE, and through the advent of federated access management.
- *Atypical individuals* (individuals whose relationship to the institution is other than staff and students, and even those categories when their relationship proceeds down a non-standard route) - These are usually handled outside the main identity management processes of an institution, both in terms of business processes and technical solutions. Ad hoc processes can lead to difficulties in accountability and security. Some particularly troublesome groups include users with NHS links, contractors, temps and employees of third party suppliers
- *Prior identity discovery* - Most institutions carry out some form of prior identity discovery, but this is usually limited to simple automated procedures or responses to users volunteering information about a previous relationship to the institution, due to the difficulty of the problem and the time it takes to carry out manual checks. The limits of the process indicate that systems that manage identities need to be able to merge identities discovered to be duplicated.
- *Virtual Organisations* - Currently, identity management for virtual organisations is carried out in an ad hoc manner. This is likely to change, with Shibboleth being singled out as a key technology for this area.