

Researcher Identifiers

Technical interoperability report



Report commissioned by JISC as part of the Digital Infrastructure Programme, on behalf of Higher Education Funding Council for England.

Report produced by Cottage Labs. Cottage Labs is a limited liability partnership, registered in Scotland with the number SO303454.

Contents

[Contents](#)

[Introduction](#)

[The form of a researcher identifier](#)

[Uniqueness](#)

[Canonicity](#)

[Semantic or Opaque](#)

[Length](#)

[URIs/URLs](#)

[Existing Identifier Strategies](#)

[HESA STAFFID \(e.g. 0311411234569\)](#)

[DOI \(e.g. doi:10.1021/ja207353x\)](#)

[UUID \(e.g. be512446-cad6-45ec-a0f3-e6cf82972188\)](#)

[ISBN \(e.g. 978-3-16-148410-0\)](#)

[The researcher profile](#)

[Overview](#)

[Technologies](#)

[Content Negotiation](#)

[Linked Open Data](#)

[APIs](#)

[Data sources](#)

[Current Technical Considerations](#)

[Existing Software of Relevance](#)

[Current Research Information System](#)

[Repository Systems](#)

[Bibliographic / Profile Systems](#)

[National Infrastructure](#)

[Existing Standards/Specifications of Relevance](#)

[Metadata Formats](#)

[Object Formats](#)

[Transport Formats](#)

[Authentication and Authorisation](#)

[APIs](#)

[Related Efforts](#)

[Case Study: Je-S System](#)

[Workflow implications](#)

[Models](#)

[Researcher owns / can claim their profile](#)

[Institution\(s\) own their researcher's profiles](#)

[Researcher identifier provider owns the profiles](#)

[Multiple Identities](#)

[Security and Privacy](#)

[Data Management Models](#)

[SWOT Analysis](#)

[Centralised](#)

[Distributed](#)

[Aggregated](#)

[Federated](#)

- [Interoperability with other systems](#)
- [Implementation issues](#)
 - [Scope and structure of the system](#)
 - [Information model](#)
 - [Identifier philosophies](#)
 - [Security and Privacy](#)
 - [Sustainability](#)
 - [Existing efforts](#)
 - [Development strategies](#)
 - [Scale of work](#)
- [Recommendations](#)
 - [Form of the Identifier](#)
 - [General Recommendations](#)
 - [Specific Recommendations](#)
 - [Concluding Remarks](#)
- [References](#)

Introduction

The need for a researcher identifier has been widely debated. This report discusses some of the technical aspects of implementing an identifier and profile system for researchers. This report is a companion piece to the report titled: '*Researcher Identifiers: Data sources report*' which provides an overview of sources of data relevant to the task of creating profiles for academic researchers in the UK.

Different people mean different things when they talk about an identifier. An identifier is purely some value (e.g. a string, number, URI, sequence of bytes...) that uniquely identifies an entity. When discussing researcher identifiers the topics of author profiles, authentication and authorisation almost inevitably come up. It is important to note that while a standard identifier system may help the development of profiles and authentication and authorisation tools, these functions do not need to be part of the identifier system. That said, in order to enable identifier discovery, it is likely that at least some 'profile data' would need to be stored by the identifier system - including name(s), date of birth, past and present institutional affiliations.

A key question is whether there will be a compulsion on the researcher to take 'ownership' of their identifier (e.g. be responsible for knowing and giving out their researcher identifier in a similar manner to their National Insurance number)? Alternatively, it may be that a researcher can remain unaware that they even have an identifier with all registration and management taken care of by institutional and system administrators, behind the scenes. Which of these models is adopted will have a large influence on the implementation of any identifier system.

There are already a large number of identifiers in use in the UK Higher Education sector (for the purposes of: HESA reporting, RCUK Joint Electronic Submissions System, institutional administrative and IT systems, digital repositories and publication systems, and so on). These systems are largely disconnected, leading to much duplication of effort and making it extremely difficult to generate a comprehensive view of an individual's interactions with the different systems.

This report discusses the following topics:

- The form of a researcher identifier
- The nature and construction of researcher profile
- The current state of technology
- Workflow challenges and implications in providing such a system
- Models and challenges for implementation

The report concludes with some recommendations as to the best approaches to building such a system.

The form of a researcher identifier

This section of the report summarises key issues to consider in developing the identifier itself (not the profile), and describes the approaches taken by a number of widely used existing identifiers.

Uniqueness

The most important factor when generating an identifier is to guarantee its uniqueness: it must not be possible for the same identifier to be assigned to two different researchers. How this can be achieved will depend on the mechanism by which identifiers are generated. Uniqueness is simplest to ensure if all identifiers are allocated by one provider, in which case it is straightforward to check against a list of previously assigned identifiers. If there are multiple identifier providers then there must be a mechanism for ensuring that the same identifier is not assigned by two different providers. The most common approaches to solving this problem are to partition the identifier space and assign each region to a single provider, or to randomly generate identifiers with sufficient entropy that any chance of a collision is reduced to an acceptable level. Alternatively providers can reserve batches of identifiers from a central authority who ensures that there is no overlap between the batches allocated to different providers.

Canonicity

It is also important that an identifier should be canonical - each researcher should ideally have only one. This is much more difficult to achieve than uniqueness, and it is unlikely to be able to be guaranteed. Uniqueness is a purely technical issue (check that the identifier has not already been assigned), while canonicity (ensuring that a researcher has just a single ID) relies on processes and workflows being correctly followed, and on disambiguation - both of which are difficult challenges. Whether by accident or design it is likely that some researchers will end up with being assigned more than one ID, and the system will need to be able to cope with this. This is discussed further in the 'Workflow Implications' section of this document.

Semantic or Opaque

Identifiers can be either semantically rich (containing 'meaning' in at least some portion - e.g. http://researchers.ac.uk/hermione_granger), or opaque (e.g. <http://researchers.ac.uk/89-69-8>). Semantic identifiers raise some clear issues in the context of a researcher identifier. Semantic information in an identifier will inevitably become outdated: researchers move between institutions; institutions can merge; researchers change their name (e.g. marriage). Alternatively data may be entered incorrectly (e.g. start date, date of birth); in which case should the identifier be changed, or left with the incorrect data?

Even if semantic information is only used when the identifier is generated, with the creators knowing and expecting that it will decay, other people may not be aware of this and will extract information from it and potentially rely on it, unaware of the problems this may cause.

There are also security and confidentiality implications regarding the use of semantic identifiers. Depending on the method used to generate the identifier, information such as a researcher's nationality or previous affiliation could be exposed.

Length

The length of an identifier can be an issue for a number of reasons. Long identifiers (especially apparently opaque identifiers such as UUIDs - e.g. 0ed8b17d-046d-4b04-a793-f34940f23a99 - and long URLs - e.g. a Google document: <https://docs.google.com/document/d/1qWiTs3pFKmFxEhQRIKdNjoGhSWCG5i3M95IKvkX4mkg/edit>) are difficult to manually enter or transcribe. Even if it is never intended that an identifier be manually recorded it is likely that it will occur, at least occasionally. Very long identifiers are likely to cause problems even with actions such as copying and pasting from emails and other documents due, for example, to issues with line-breaks.

URIs/URLs

A further question is whether an identifier should be URI/URL [1] based. A URI (Uniform Resource Identifier) is a string of characters used to identify a name or a resource on the Internet. A URL (Uniform Resource Locator) is a special class of URI as well as identifying a resource, provides a means of locating it on the network (e.g. its HTTP address). In order to inter-operate with the growing Semantic Web it is essential that an identifier can be represented as a URI, however whether this should be an HTTP URL is open to debate.

If an HTTP URL is used as an identifier (e.g. <http://id.researchers.ac.uk/abc123>) then it will be expected to be actionable, and resolve to a page providing information about the subject of the identifier. This gives a great deal of power and control to the owner of the domain under which the identifier is provided (in this case researchers.ac.uk).

Rather than a single domain, a distributed system could permit any URL to be used as an identifier, in a similar manner to OpenID [2]. Researchers could have identifiers such as <http://people.ucl.ac.uk/xyz789>, <http://ids.ox.ac.uk/abc123> and even <http://mypersonalsite.co.uk/me>. This approach raises other issues. If institutional addresses are used, what happens when a researcher changes institution, or leaves academia?

Use of HTTP URLs also raises issues of persistence. What happens if a domain disappears? This is very likely to occur if personal domains can be used, and it is not unheard of for institutions to re-brand or merge.

DOIs can be expressed using “info:” URIs (e.g. <info:doi/10.1000/182>) which are non-actionable, and under no organisation’s control. The International DOI Foundation provide a DOI resolver (<http://dx.doi.org/>) which acts as an HTTP proxy redirecting users to the subject of the DOI.

Example:

Basic format	doi:10.1021/ci200309j
info URI	info:doi/10.1021/ci200309j
HTTP Proxy URL	http://dx.doi.org/10.1021/ci200309j
Result of redirect	http://pubs.acs.org/doi/abs/10.1021/ci200309j

Existing Identifier Strategies

HESA STAFFID (e.g. 0311411234569)

HESA STAFFIDs [3] are assigned by a researcher’s institution. The ID is prefixed by the year in which

the staff member entered the institution, the ID number of the institution and a six digit reference number assigned by the institution. The year and institution ID act to partition the identifier-space (thus incorporating some semantics), leaving the institution responsible for ensuring the uniqueness of the reference number between staff joining each year.

DOI (e.g. doi:10.1021/ja207353x)

DOIs are formed of two components - a prefix and a suffix. The prefix is specified by the DOI registration agency (e.g. CrossRef) and identifies the organisation that registered the DOI, and the suffix is chosen by that organisation and identifies the specific digital object. While DOIs may initially appear opaque, they can contain substantial semantic information: the prefix identifies the organisation registering the DOI, and the suffix often contains information that identifies the journal, and in some cases even includes the volume, issue and page numbers of an article.

UUID (e.g. be512446-cad6-45ec-a0f3-e6cf82972188)

There are a number of versions of UUID, which provide uniqueness in different ways. Version 1 UUIDs are based on the MAC address of the machine generating the UUID and the current timestamp. Version 4 UUIDs are generated using random numbers, and are 36 hexadecimal characters long. Assuming that the random number source provides sufficient entropy, the size (128 bits) of the UUID means that there is minimal chance of producing a duplicate.

ISBN (e.g. 978-3-16-148410-0)

The ISBN identifier-space is partitioned into 'blocks' of sequential identifier numbers. A national registration agency (Nielsen Book Services [4], in the UK) assigns blocks of ISBNs to a publisher (for a fee) to allocate to publications as they choose. Once a publisher has exhausted their block of ISBNs they are assigned a new block to use.

The researcher profile

Overview

One of the primary benefits of a unique researcher identifier would be to support the construction of researcher profiles. There is currently no such thing as researcher profile - only multiple fragmented sources of information. At present the most successful of these, from the point of view of researcher uptake and reuse, are those managed and driven by the researchers themselves - i.e university homepages, or profile pages held within services run by commercial organisations like Facebook [5] and LinkedIn [6].

There are a wide variety of data sources with information on researchers, and the existence of a common identifier system would vastly simplify their aggregation (see the Introduction to the *Researcher Identifiers: Data sources* report). Such profiles would be of wide use to institutions, funders, national agencies such as HESA, and especially researchers themselves.

Researcher profiles could be quite separate to a system for administering researcher identifiers - an identifier resolver service could point anywhere, and that it points at a profile service would be very useful, but not an absolute requirement. Similarly the data for constructing a researcher profile need not be stored in the same system that provides the identifier. Linking the two (or at least launching them simultaneously) could help to spread awareness and boost uptake of an identifier system. It is also likely that an identifier system will need to store at least some of the information that would be included in a profile, in order to support discovery mechanisms, so it may make sense to integrate identifiers and profiles together.

A researcher profile needs to be able to support a wide variety of information: researcher name(s), contact details and biographic information; current and past affiliations; research outputs; and possibly links into other systems with information about the researcher. A wider range of research outputs need to be recordable: traditional publications such as journal articles, chapters and whole books, along with other types of research output such as patents, press reports, data sets, blog posts, software, educational resources and artistic works (paintings, videos, manuscripts...). It is impossible to create an exhaustive list of the types of data that will need to be recorded, so a mechanism of distributed extensibility (allowing new types of information to be added) will be essential.

Technologies

All of the technologies to build a researcher profile already exist - there are many researcher identifiers / profiles already in existence - e.g. arXiv [7], scopus [8], MS Academic Research [9], University of Southampton ECS's academic's profiles [10] - though none are comprehensive. Many current technical efforts relate only to interoperability within closed systems - there is no consensus or generalized standards for the sector.

It is important that any new system should work with existing standards and technologies, rather than creating new ones. Widely used standards around identity and profile exchange and management include Dublin Core [11], FOAF [12], vCard [13], PRISM [14], BIBO [15]. Common formats are JSON [16], XML [17], RDF [18] (XML, N3).

Content Negotiation

Content negotiation provides a mechanism for web servers to provide alternative representations of a resource from the same URL, depending on the preferences a client expresses. A common use is for a server to return either a PNG or a GIF version of an image, depending on which the client prefers. Use of content negotiation would enable a researcher profile system to return HTML pages to a web browser, and machine readable (e.g. JSON, RDF/XML...) to software agents.

Linked Open Data

Linked (Open) Data [19] is one of the ways in which the Semantic Web is developing. It provides a mechanism for publishing structured data across the web in a way that it can become interlinked, forming a more useful whole. The basic principles behind linked data are to use HTTP URIs to identify 'things', and to provide useful machine readable information about those 'things', along with links to other 'things' when the URI is dereferenced. This means that given an initial identifier software agents can download information about the subject of that identifier, and retrieve links to further information and other related 'things'. Use of content negotiation is also valuable here, as resources can be presented as web pages to humans and the appropriate format for machine clients.

One of the major challenges facing the Semantic Web is uniquely identifying individuals. A researcher identifier that is, or can be represented by, a URI could be used to address a number of challenges in this area, creating efficiencies and globalization of researcher data. However, this may generate significant privacy issues.

APIs

If a researcher profile supports content negotiation and returns linked open data, then it can function as an API, allowing other services to connect to and build upon the profile system. For example, arXiv provide a javascript gadget [20] that enables authors to embed a 'recent publications' box in their person websites, using the arXiv API.

The same privacy and security considerations that arise from the profile apply to the API: what (if anything) should be public? Who should have access? Who controls access? It is likely that some researchers will be happy for most of their profile to be public, whilst others (particularly those engaged in sensitive areas, such as animal research) would require all of their information to remain private. A further consideration is whether the API should be read-only, or if it should permit properly authenticated users to update the profile information?

The API could provide views into subsets of a researcher's profile - e.g. publication lists, or grant applications - however, it is likely that the greatest value to institutions and the likes of HESA would come from additional APIs providing query functionality and views into aggregations of profiles, rather than looking at one individual's.

Data sources

How a researcher profile is generated will depend on the data management model adopted by the system. The profile could be generated from a single data store, holding all the relevant data for that researcher as per *Figure 1*.

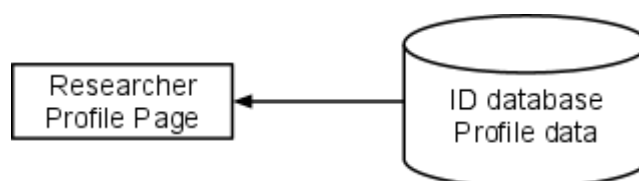


Figure 1: researcher profile generated from single data store

Alternatively, a profile could be generated by dynamically accessing a number of different data sources, and aggregating the results, as per *Figure 2*.

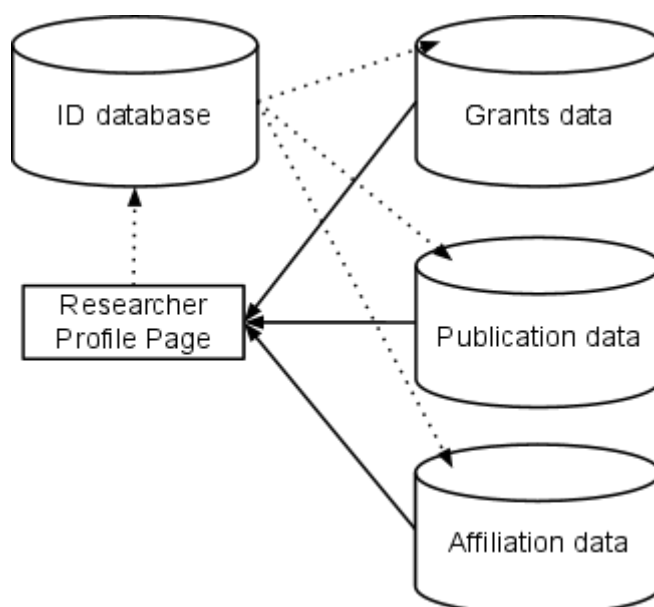


Figure 2: researcher profile generated dynamically from multiple datasources

A centralised data store will require a flexible data model, and may be complex to keep up-to-date, while there are issues of discovery in the distributed system. How does the profile know where to find information on a particular researcher? Are all the data sources indexed by the researcher identifier database, or will there a mapping to their key systems be needed? Some combination of these models is also feasible; for instance, a central store may hold 'core' details of a researcher, such as name, contact details and affiliations, and further information such as lists of publications and other research outputs may be retrieved through dynamic queries.

Current Technical Considerations

There are many existing software systems, specifications and standards relevant to a researcher identifier / profile system. Software exists providing many of the functions of such a system, and is important that lessons can be learned from their experiences. There are also many systems that could provide information to, and use information from a researcher identifier / profile system.

Existing Software of Relevance

It is important that mechanisms for interacting with software is considered, and commonly used specifications and standards supported. Unfortunately there are many such specifications and standards, and a lack of consensus about their use. Illustrative examples are given below.

Current Research Information System

Current Research Information Systems (CRIS) are tools that provide institutions with a picture of their research interests, funding and other related information. They can provide local researcher pages from the data they contain, and may expose their data through APIs, but these are generally proprietary.

System name	Description
Symplectic Elements [21]	Elements is a proprietary CRIS system developed by Symplectic in the UK
Pure [22]	Pure is a proprietary CRIS system developed by Atira in the Netherlands
Converis [23]	Converis is a proprietary CRIS system developed by AVEDAS in Germany

Repository Systems

Institutional repositories collect and preserve an institution's research outputs, increasing the visibility of the institution's research. They can provide a rich source of data about an institution's research outputs (and in particular full-text content), however the degree of usage and quality of metadata they contains varies widely.

System name	Description
EPrints [24]	EPrints is an open source digital repository system written in Perl, originally developed at the University of Southampton
DSpace [25]	DSpace is an open source digital repository system written in Java, originally developed at MIT and HP Labs
Fedora Commons [26]	Fedora Commons is an open source digital asset management system written in Java, originally developed by Cornell University and the University of Virginia

Bibliographic / Profile Systems

A number of bibliographic / profile systems are available. These provide standard means for institutions to implement local profiles for their academics and other researchers.

System name	Description
BibApp [27]	BibApp is an open source Institutional publications / CV system.
VIVO [28]	VIVO is an open source semantic web application for publishing and exploring researcher interests and activities.
Harvard Catalyst [29]	A publications/researcher profile navigator developed at Harvard

National Infrastructure

System name	Description
HESA [30]	HESA collect, analyse and distribute quantitative information about UK higher education, including staff and students.
Je-S [31]	The Je-S System is used by many UK research councils and other funding bodies to provide their communities with electronic grant services.

Existing Standards/Specifications of Relevance

There are many existing standards and specifications relevant to a researcher identifier / profile system:

Metadata Formats

A number of ontologies / vocabularies are widely used to describe data relevant to a researcher profile. In order to provide the greatest possible scope for interoperability, these existing formats should be employed so far as possible.

Vocabulary	Description
Dublin Core [11]	For describing common metadata terms
FOAF [12]	For describing people and organisations
PRISM [14]	For describing bibliographic information
BIBO [15]	For describing bibliographic information

Object Formats

Format	Description
OAI-ORE [32]	Provides an extensible data model for describing aggregations of objects, expressible in a variety of formats
Atom Syndication Format [33]	An XML format for providing feeds of web objects. The Atom format is widely used to provide feeds of blog posts, and is increasingly being used to provide feeds of other types of resources.
METS [34]	Provides is a widely used schema for encoding metadata about objects in digital libraries.
DIDL [35]	Fulfils a similar role specifically for multimedia content, and is part of the MPEG specifications.

Transport Formats

Format	Description
AtomPub [36] / SWORD [37]	Provide a mechanism for communicating digital artifacts between scholarly systems. They are most widely used to provide deposit mechanisms to institutional repositories, but could be applied to other tasks such as transferring bibliographic data.
OAI-PMH [38]	Provides a widely used protocol for harvesting metadata from a system. It is likely the a researcher profile system will need to use OAI-PMH to collect information about research outputs, for example harvesting the content of institutional repositories. An OAI-PMH server could also be offered as part of the system's API.

Authentication and Authorisation

Protocol	Description
Shibboleth [39] / Athens [40]	The Shibboleth and Athens systems to provide federated authentication / single sign on through one's home institution.
OpenId [2]	OpenID provides a similar mechanism for the wider web, enabling the use of many systems (e.g. Google, Facebook, Twitter) as an identity to sign into other systems.
OAuth [41]	The OAuth specifications enable a user to delegate some or all of their privileges on a system to another user, without having to share their credentials, and in a revocable manner.

APIs

The greatest benefits of a researcher profiles system are likely to be realised through the development and integration of other services that build on the data in the researcher profiles. This will require the profile system to expose an API.

API	Description
OpenSearch [42]	Provides mechanisms for sharing search APIs and results.
OpenURL [43]	Helps to locate copies of resources that a user is allowed to access.
SPARQL [44]	While it is not safe to provide direct query access to an SQL database, if an RDF triplestore provides the data back-end then a SPARQL endpoint can be exposed, though this may not be able to support fine-grained privacy and Authentication and Authorisation controls.

Related Efforts

There are a variety of ongoing efforts in the area of researcher identifiers and profiles. A number of countries have implemented some form of researcher identifier / profile:

Country	System
Netherlands	Digital Author Identifier [45]
Norway	CRISTIN [46]
Australia	People Australia [47]
New Zealand	NZETC EATS [48]
Japan	Researcher Name Resolver [49]
Germany	DissOnline [50]
Brazil	LATTES [51]
UK	HESA IDs [3]

There are also a number of existing efforts with international scope:

System
Elsevier SCOPUS (commercial, closed-access) [8]
Thomson Reuters' Researcher ID (commercial, closed-access) [52]
Microsoft Academic Search (commercial, open-access) [9]
Google Scholar Citations (commercial, open-access) [53]

AuthorClaim (non-commercial, open-access) [54]
--

ORCID (not yet launched, planned open-access) [55]
--

Of these systems, ORCID looks to offer the most promising, largely due to the scope of its support (over 250 participating organisations including many major scholarly publishers), however the live system is not yet available so it impossible to judge whether its potential will be realised.

It is also worth noting the increasing interest in this area from global organisations such as Microsoft and Google. Microsoft's academic search platform (launched earlier this year) includes researcher publication profiles (e.g. <http://academic.research.microsoft.com/Author/10744969>), and encourages researchers to submit corrections to their own (and other's) data. The platform provides a limited API giving some opportunity to build other services on top of it. Google are also in the process of adding 'Citation Profiles' to their Google Scholar search results (e.g. <http://scholar.google.com/citations?user=cKBTIzIAAAAJ>). These are created from a mixture of algorithmically generated matches, and researchers claiming and administering their profiles.

Case Study: Je-S System

The RCUK Je-S System holds information on names and addresses of staff involved in successful grants, unsuccessful proposals and reviewing, and some patchy information on areas of expertise. The system has its own identifier, called the CDR-ID. The CRD-ID is complementary to HESA's STAFFID, and includes some researchers who work outside of HEIs. RCUK are supportive of the aims of a standardised researcher identifier / profile system, with a particular interest in disambiguation of research outputs. Some information from Je-S could potentially be released into such a system, however there are significant data protection issues that remain to be resolved.

A key feature of a researcher identifier / profile system that J-eS could take advantage of is the ability to give RCUK access to information to enable performance review or evaluation of bidders prior to making an award. This would both enhance the effectiveness of the awards process and also ensure that the researcher identifier / profile system was widely used by researchers, who would have a vested interest in ensuring the information was complete and up to date.

Je-S is used by AHRC, BBSRC, EPSRC, ESRC, MRC, NERC and STFC (formerly CCLRC and PPARC), as well as the Technology Strategy Board (TSB) and Energy Technologies Institute (ETI) , to provide their communities with electronic grant services.

Workflow implications

In this section we consider the key workflows for interacting with a researcher identifier system. The primary consideration must be who is responsible for a researcher's identifier and the associated information? Is it the researcher themselves, their current institution(s), or some other body? A further important question how much of the processes can be automated?

A researcher identifier/profile system will need to support many scenarios, for example: joining an institution, leaving an institution (and HE?), moving institution, providing access to your data (e.g. to a funding body), recording research outputs, modifying or updating data, discovery of profiles (e.g. upon staff joining new institution, from an administrator's perspective). However, these can be summarised as:

- Creating an identifier
- Finding an existing identifier
- Updating profile information associated with an identifier
- Providing (delegating) access to profile information associated with an identifier

The key issues that must be considered are, who owns/controls/is responsible for the data? Who is authorised to access the data, and who decides/grants this access?

It is likely that some researchers will want the majority of their profile information to be public, while others will want their access to their profiles to be highly restricted, possibly for security reasons (e.g. those involved in animal research), or due to concerns about competition. It is essential that information is accurate (with a mechanism to fix inaccuracies). If a profile is viewed as authoritative then its data will be used to generate metrics and make funding decisions, and any errors could have significant consequences for a researcher's career.

Models

Researcher owns / can claim their profile

- The researcher can control content and access levels
- Others (e.g. institutions, funding agencies) may be able to make assertions about the researcher (e.g. Prof. Snape is a lecturer at Hogwarts College); the researcher will be notified and can accept or dispute the assertion
- The researcher may be able to delegate control of their profile to others (e.g. institutional administrators)

Institution(s) own their researcher's profiles

- The institution is responsible for discovering / creating identifiers for their staff
- It may be difficult to decide on correct privacy levels without researchers' input
- How are researchers with multiple affiliations supported?
- Who becomes responsible for the profiles when a researcher leaves?

Researcher identifier provider owns the profiles

- This will still require input from researchers / institutions to acquire data
- It must be highly trusted not to act purely in self-interest

Multiple Identities

While the majority of researchers will have a single identifier and identity, some researchers will almost certainly acquire more than one identity as per *Figure 3*.

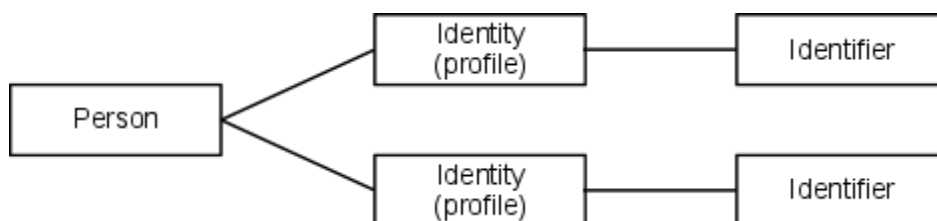


Figure 3: researcher with multiple identities and profiles

When this happens the identities will need to be merged. Depending on the information stored in each profile, some manual reconciliation is likely to be needed. The obvious question is what should be done with the identifiers? The researcher identifier system is unlikely to be able to know everywhere that the identifiers are used, so the now redundant identifiers cannot simply be deleted - it is probably being used somewhere. Instead, there will be a situation where a number of identifiers are now associated with a single identity, as per *Figure 4*.

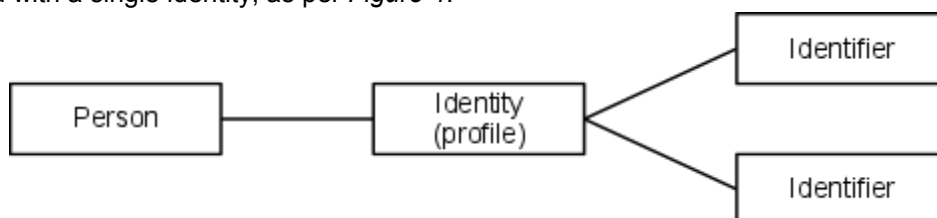


Figure 4: a single researcher profile linked to by two identifiers


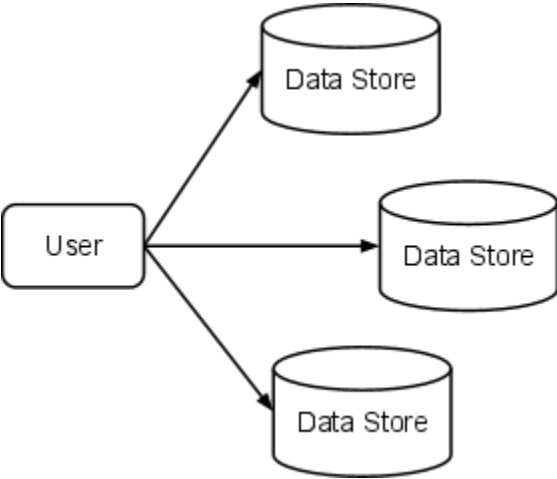
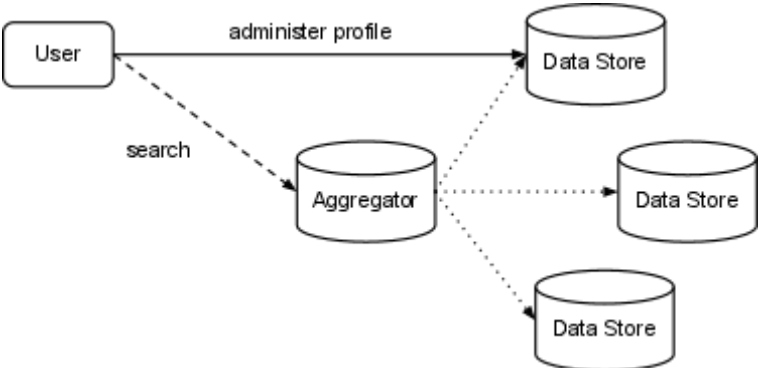
It may also be the case that some researchers need more than one identity. For example a researcher who spends some of their time / career engaged in controversial (e.g. animal research) or sensitive (e.g. national security related) activities may want one identity for aspects of their work that can be made public, and a separate identity for the more sensitive aspects.

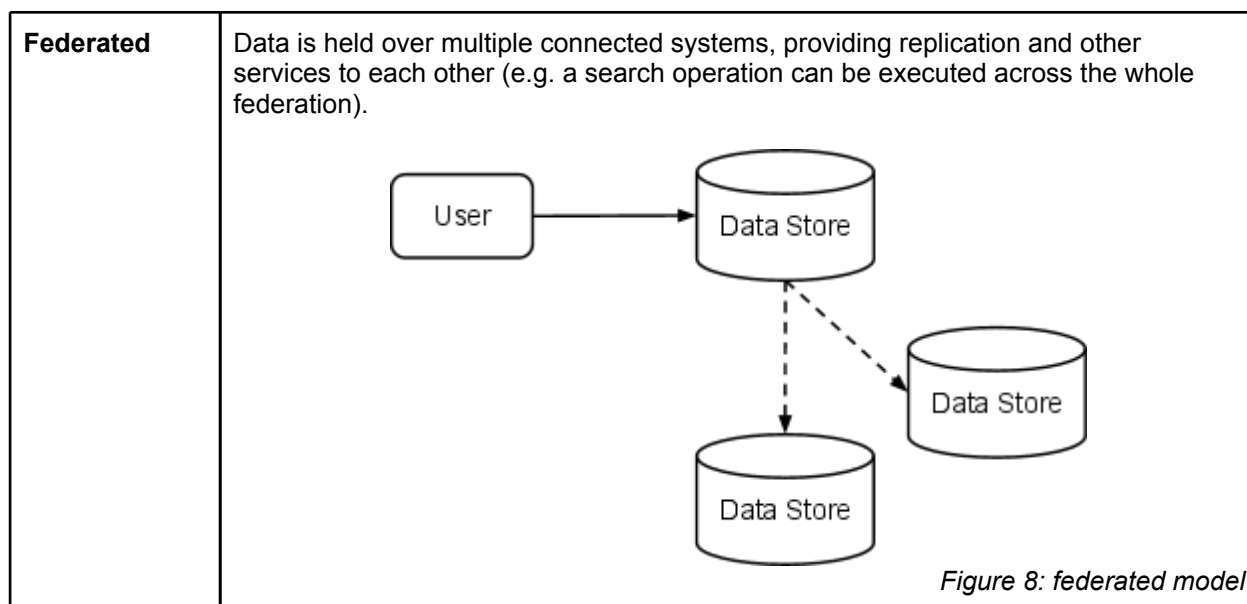
Security and Privacy

Strong security and privacy are essential for a researcher identifier / profile system. All the necessary security technologies exist, and there are software frameworks to support their use, however good security is not straightforward to understand or implement, and must be considered throughout the design process. A critical issue will be to ensure respect for privacy settings in downstream systems. This is an issue of trust, and does not have a purely technical solution.

Data Management Models

There are a variety of data management models that could be applied to researcher identifier and/or profile systems. Each has strengths and weaknesses, and impacts on interoperability with other systems.

Centralised	<p>Users interact with a single system that holds all data. There could be local user interfaces, but all operations must be validated against this system.</p>  <p><i>Figure 5: centralised model</i></p>
Distributed	<p>Data is held in multiple systems, with little communication between them. Operations must be executed on each system individually.</p>  <p><i>Figure 6: distributed model</i></p>
Aggregated	<p>Data is held in multiple systems, but with one or more aggregators providing (possibly stale) views over the whole data set.</p>  <p><i>Figure 7: aggregated model</i></p>



SWOT Analysis

This section outlines the relative benefits and drawbacks of data management models describe above.

Centralised

Strengths	Weaknesses
<ul style="list-style-type: none"> Consistency of data Completeness of data Easy to get started with Easy to maintain Easier to do global operations (e.g. disambiguation, search, etc) 	<ul style="list-style-type: none"> Single point of failure Lack of local control A reluctance by institutions to rely on single authority Large single cost Overall sustainability
Opportunities	Threats
<ul style="list-style-type: none"> Standards enforcement within the sector Changes in policy can be enforced Simplicity in implementation and management Simpler to build services on Lower overall costs Valuable to national services, etc 	<ul style="list-style-type: none"> Competition from alternative models Security Data protection Policy continuity may threaten sustainability

Distributed

Strengths	Weaknesses
<ul style="list-style-type: none"> • No single point of failure • Organisations can have local control • Individual nodes can respond rapidly to changing needs 	<ul style="list-style-type: none"> • Each node is isolated • Maintenance requires duplicated work around the network • Lack of authority control • Overall more complex • Less usable from an individual perspective
Opportunities	Threats
<ul style="list-style-type: none"> • Organisations can customise locally • Enhanced buy-in • Distributing the effort and technical skill of set up. 	<ul style="list-style-type: none"> • Divergence from standards • Lack of effort at an organisational level • Variation in technical skills • Higher overall cost • Potentially patchy service

Aggregated

Strengths	Weaknesses
<ul style="list-style-type: none"> • No single point of failure • Organisations can have local control • Individual nodes can respond rapidly to changing needs • Central point of contact for users • Easier to do global operations (e.g. disambiguation, search, etc) • Potentially complete (although not guaranteed) 	<ul style="list-style-type: none"> • Potential duplication of costs • Laggy (aggregation takes place after-the-fact) • Not an authority, although has the potential • Overall complexity • Maintenance requires duplicated work around the network • Potential inconsistency
Opportunities	Threats
<ul style="list-style-type: none"> • Organisations can customise locally • Enhanced buy-in • Distributing the effort and technical skill of set up. • Flexibility to deal with new systems/sources • Possible to enforce standards to some degree • Simpler to build services on 	<ul style="list-style-type: none"> • High turnover of information • Divergence from standards • Higher overall cost • Security • Data protection • Technically diverse environment requires a lot of work to aggregate from

Federated

Strengths	Weaknesses
<ul style="list-style-type: none"> • No single point of failure • Organisations can have local control • Individual nodes can respond rapidly to changing needs • Resilient as a network • More accessible from a user point of view 	<ul style="list-style-type: none"> • Potential inconsistency • No central point of contact (although Federated model would support aggregators implicitly) • Potential duplication of costs • Potentially laggy (depending on communications mechanism) • Maintenance requires duplicated work around the network • Lack of authority control • Overall complexity
Opportunities	Threats
<ul style="list-style-type: none"> • Fail-over possible • Shared services are viable • Organisations can customise locally • Enhanced buy-in • Distributing the effort and technical skill of set up. • Flexibility to deal with new systems/sources • Possible to enforce standards to some degree 	<ul style="list-style-type: none"> • Higher overall cost • Divergence from standards • Technically diverse environment would generate a lot of work • Lack of effort at an organisational level • Variation in technical skills

Interoperability with other systems

It is essential that researcher identifiers and profiles can interact with other (scholarly) systems. Where possible this can be facilitated by systems storing researcher identifiers as foreign keys (references to other identifier systems) for their records, however many systems of interest may not be aware of the researcher identifiers' existence. In these cases it will be necessary to establish mappings between their identifiers and the researcher identifier. How these links, which will not necessarily be one-to-one, will be developed and maintained must be considered.

Implementation issues

We do not foresee any insurmountable technical barriers to the provision of a researcher identifier system, however there are a number of critical issues that any implementation of a researcher identifier needs to address:

Scope and structure of the system

The first questions that need to be addressed are the scope and structure of the system. The scope of the system needs to be clear: provision of identifiers and profiles are potentially different tasks; while these *could* be the same system, they don't necessarily need to be. Similarly, the choice of data management model will have a large impact on the implementation of the system.

Information model

The choice of information model will be strongly affected by the scope of the system. An important consideration for a profile service is whether there should be any mechanism for extensibility - will the model be strictly defined, or should it be able to support the addition of arbitrary information. The ability to disambiguate and merge identifiers/profiles will be essential; some aspects can be automated, but in the end manual input will be required.

Identifier philosophies

Any discussion about identifiers ends up in (often circular) argument! Failure to accept that there is no 'perfect' answer will lead to endless debate. Decisions need to be pragmatic; someone has to take responsibility and ensure decisions are made. Something that works is better than a hypothetical perfect system that never materialises.

Security and Privacy

This is the main technical issue. Security and trust are critical issues for any system handling personal (and potentially confidential) information. These are hard to get right, especially in a distributed environment, and will require considerable time to develop and thoroughly test. To drive the technical development clear policy is essential, which is also complex and time consuming to develop.

Sustainability

The sustainability of a researcher identifier system must be addressed. Besides initial development, considerable effort will be required for continued maintenance, extension and local customisation. A single national service is likely to have lower overall costs than many local institution managed systems. This cannot simply be a short-term 'project'.

In order to build trust and gain wider acceptance a clear commitment for the long-term future of the system will be needed. It is important that any project has stakeholder buy-in to ensure success. This will require initial sponsors to put considerable resources into advocacy and outreach activities, and proving the use cases for the service, before a sustainable business model can be developed.

Existing efforts

A number of current and upcoming researcher identifier / profile systems were described previously. It is likely that they will be able to support at least some aspects of any required system. Adapting

or building upon these efforts may prove substantially more cost effective than developing a new system from scratch, and is likely to help improve interoperability and standards compliance. Of those projects (listed previously) ORCID stands out as the one with the greatest cross sector buy-in and potential for high uptake.

Development strategies

Choice of software development process will clearly influence the cost and timescale of implementing a system, but can also affect the acceptability and community buy-in of the finished system. An open source, agile development strategy will help to build trust and community engagement from the start of the project, while enabling the developers to support refinements and changes to requirements that will inevitably emerge as implementation progresses.

Scale of work

Centralised	A single larger development team. One production environment is all that is necessary, so it can be tightly controlled, reducing overall complexity. Individual organisations may then create local views/interfaces via APIs.
Distributed	This would likely consist of multiple smaller teams. Overall this is likely to require more resources than centralised approach. Because of the need to support heterogeneous environments, there is a significant increase in complexity.
Aggregated	Similar to Distributed in terms of resources, but with additional development of initial (and perhaps canonical) aggregator systems, with some possible work on standards efforts.
Federated	Multiple smaller teams requiring high degree of coordination. Significant project management and standards efforts would be required, so this would be the most costly approach.

Recommendations

Building on the discussions in this report we can make the following recommendations.

Form of the Identifier

The DOI system inspires many of our recommendations about the form of the identifier.

- It should be opaque, to avoid issues with outdated or inaccurate semantic information
- It should have a well-defined structure such that a string can be recognised as a likely researcher identifier without any context (e.g. 1-234-56-7)
- There should be a URI representation (e.g. info:rid/1-234-56-7), to support semantic web applications
- It should not be an HTTP URL as this gives too much authority to the domain owner
- There should be a standard resolver system (e.g. http://researcher.ac.uk/1-234-56-7)
- Allow scope for other resolvers, possibly only supporting a subset of IDs, such as those for researchers at a single institution (e.g. http://people.hogwarts.ac.uk/1-234-56-7)

Development Recommendations

- Maximise community buy-in, and minimise disruption from shifting requirements
- Development efforts should be open source, agile, incremental/phased
- Large effort must be committed to advocacy and community development
- Detailed specification work is needed, with input from a broad range of stakeholders

General Recommendations

There should be a comprehensive researcher identifier / profile system. The current variety of fragmented systems are not meeting the needs of the sector. A researcher identifier / profile system should:

- It should be offered at a national level, not rely on individual institutions
- An effort must engage technical people from many interested parties
- Although national in scope, it must have international range
- It should ensure support researchers from outside traditional HEIs
- Embrace current standards
- It should support semantic web / linked data applications
- There should be an open API enabling other applications to build on the service

Specific Recommendations

- To be able to provide a 'researcher CV' service supporting the variety of modern research outputs
- Set up a national service - e.g. researchers.ac.uk as the canonical resolution and possibly profile service
- Researchers must be able to take charge of their profiles, with an option to delegate to their institution(s)
- There should be fine-grained privacy controls
- Security must be done right from the start
- Monitor the progress of the ORCID project, with a view to utilising / inter-operating where possible

Concluding Remarks

As is so often the case, the most complex issues are political rather than technical.

Whilst federation and other suitable technologies exist, each requires a compromise on certain functionalities; deciding which to abandon will be a matter of managing stakeholder expectations. The effort required to address profile privacy and security should not be underestimated, and identification of which data can reasonably be expected to remain private will be key. No matter what technical solution is employed, there is an intrinsic trade-off between consistency, availability and partitioning.

References

1. URIs, URLs, and URNs: Clarifications and Recommendations 1.0 - <http://www.w3.org/TR/uri-clarification/>
2. OpenID - <http://openid.net/developers/specs/>
3. HESA STAFFID - http://www.hesa.ac.uk/index.php/component/option.com_collns/task.show_manuals/Itemid.233/r.08025/f.003/
4. Nielson UK ISBN Agency - <http://www.isbn.nielsenbook.co.uk/controller.php?page=121>
5. Facebook - <http://www.facebook.com/>
6. LinkedIn - <http://www.linkedin.com/>
7. arXiv Author Identifiers - http://arxiv.org/help/author_identifiers
8. Scopus Author Identifier - http://help.scopus.com/robo/projects/schelp/h_auteursrch_intro.htm
9. Microsoft Academic Search - <http://academic.research.microsoft.com/>
10. University of Southampton ECS Academic Profiles - <http://www.ecs.soton.ac.uk/about/researcher-profiles.php>
11. The Dublin Core Metadata Initiative - <http://dublincore.org/>
12. FOAF Vocabulary Specification - <http://xmlns.com/foaf/spec/>
13. Personal Data Interchange - vCard and vCalendar - <http://www.imc.org/pdi/>
14. PRISM Specifications - <http://www.prismstandard.org/specifications/>
15. The Bibliographic Ontology (BIBO) - <http://bibliontology.com/>
16. JavaScript Object Notation (JSON) - <http://www.json.org/>
17. W3C - Extensible Markup Language (XML) - <http://www.w3.org/XML/>
18. W3C - Resource Description Framework (RDF) - <http://www.w3.org/RDF/>
19. W3C - Linked Data - <http://www.w3.org/standards/semanticweb/data>
20. Using the arXiv myarticles widget - <http://arxiv.org/help/myarticles>
21. Symplectic Elements - research management system - <http://www.symplectic.co.uk/products/publications.html>
22. Atira Pure Current Research Information System - <http://atira.dk/en/pure/>
23. Avedas CONVERIS Current Research Information System - <http://www.avedas.com/en/converis.html>
24. EPrints - <http://www.eprints.org/>
25. DSpace - <http://www.dspace.org/>
26. Fedora Commons - <http://fedora-commons.org/>
27. BibApp - <http://bibapp.org/>
28. VIVO - <http://vivo.sourceforge.net/>
29. Harvard Catalyst Profiles - <http://catalyst.harvard.edu/spotlights/profiles.html>
30. The Higher Education Statistics Agency (HESA) - <http://www.hesa.ac.uk/>
31. RCUK Je-S System - <https://je-s.rcuk.ac.uk/>
32. Open Archives Initiative Object Reuse and Exchange (OAI-ORE) - <http://www.openarchives.org/ore/>
33. The Atom Syndication Format - <http://tools.ietf.org/html/rfc4287>
34. Metadata Enhancement and Transmission Standard (METS) - <http://www.loc.gov/standards/mets/>
35. Digital Item Declaration Language (DIDL) - <http://xml.coverpages.org/mpeg21-didl.html>
36. The Atom Publishing Protocol - <http://bitworking.org/projects/atom/rfc5023.html>
37. SWORD - <http://swordapp.org/>
38. Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) - <http://www.openarchives.org/pmh/>
39. Shibboleth - <http://shibboleth.internet2.edu/>

40. Athens - <http://www.openathens.net/>
41. OAuth - <http://oauth.net/>
42. OpenSearch - <http://www.opensearch.org/>
43. OpenURL - <http://en.wikipedia.org/wiki/OpenURL>
44. SPARQL Protocol for RDF - <http://www.w3.org/TR/rdf-sparql-protocol/>
45. Digital Author Identifier - SURFfoundation - Netherlands - <http://www.surffoundation.nl/en/themas/openonderzoek/infrastructuur/Pages/digitalauthoridentifiervai.aspx>
46. CRISTIN - Norway - <http://www.cristin.no/as/WebObjects/cristin>
47. People Australia - Australia - <http://www.nla.gov.au/initiatives/peopleaustralia/>
48. NZETC EATS - New Zealand - <http://www.nzetc.org/>
49. Researcher Name Resolver - Japan - <http://rns.nii.ac.jp/>
50. DissOnline - Germany - <http://www.dissonline.de/>
51. LATTES- Brazil - <http://lattes.cnpq.br/>
52. Thomson Reuters Researcher ID - <http://www.researcherid.com/>
53. Google Scholar Citations - <http://scholar.google.com/intl/en/scholar/citations.html>
54. AuthorClaim- <http://authorclaim.org/>
55. ORCID - <http://orcid.org/>