cottage labs

# Researcher Identifiers

## Data sources report

JISC

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

# Introduction

This report provides an overview of sources of data relevant to the task of creating profiles for academic researchers in the UK. A researcher profile offers many inherent benefits at both the individual, institutional and national level and this report highlights some of the salient systems in use. This report is a companion piece to the report titled: *'Researcher Identifiers: Technical Interoperability'* which discusses some of the technical aspects of implementing an identifier and profile system for researchers.

At present the concept of a researcher profile is ill-defined and this report aims to go some way towards answering the question of what constitutes such a system. It includes an overview of data sources, listing the main systems that hold data about researchers in the UK, and detailed case studies of eight illustrative data sources.

There are significant privacy and trust issues associated with making personal information available via a profile and this report aims to highlight some of the key concerns in relation to the sources of data identified. In particular, any system that deals with private personal information will have potential legal implications.

## What question does a Researcher Profile answer?

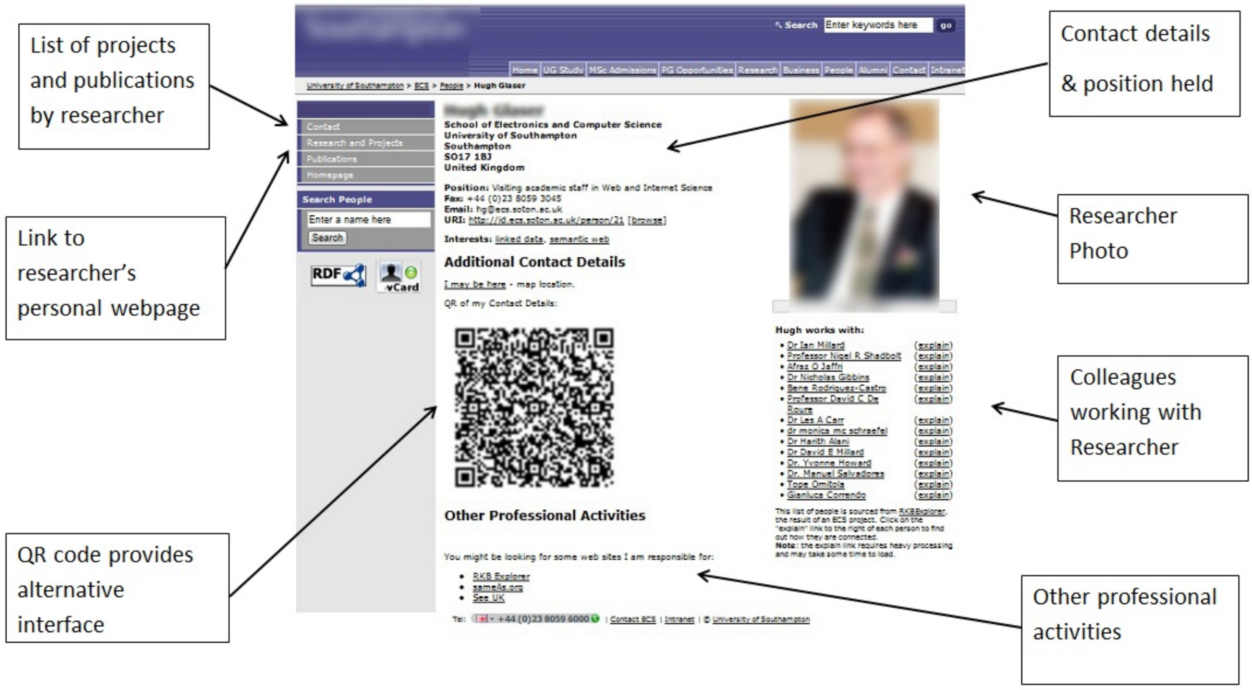
Before considering the form of a researcher profile it is important to ask what problems such a profile aims to solve. This report identifies three key aspects that drive demand for profiles within academia:

- **Work output** - A primary demand from a researcher perspective is undoubtedly the desire to present data highlighting their research achievements. This can be valuable in obtaining funding, applying for new positions, attracting co-workers or even students to an institution. Types of data could include - published papers, conferences attended, etc. This could also be important for both institutional and national level reporting such as Staff Performance Reviews or the Rsesearch Excellence Framework.

- **Career history** - A researcher will also frequently need to present data illustrating their career history, academic CV or biography. Again this is driven by the requirement to attract funding, apply for new positions, and attract co-workers or even students to an institution. Types of data could include - previous positions held, awards gained, qualifications, etc. There is also interest from bodies, such as HESA [1], as a reporting/analytical device to understand progression of professionals.

- **Disambiguation -** The ability to distinguish between different researchers with the same name via a unique identifier is a key driver from an institutional or national perspective but less so from a researcher viewpoint.

For a given use case a researcher profile will generally be a webpage containing the relevant information about the researcher and their activities. This will normally include: biographical information, contact details, photo, research interests, publications and any other aspects that the researcher feels are relevant. *Figure 1* shows an example researcher home page from the University of Southampton. It demonstrates some of the key components of a generic researcher profile including researcher biographic and contact details as well as publication information.



University of Southampton > ECS > People > Hugh Glaser

School of Electronics and Computer Science
University of Southampton
Southampton
SO17 1BJ
United Kingdom

**Position:** Visiting academic staff in Web and Internet Science
**Fax:** +44 (0)23 8059 3045
**Email:** hg@ecs.soton.ac.uk
**URI:** http://id.ecs.soton.ac.uk/person/21 [browse]

**Interests:** linked data, semantic web

**Additional Contact Details**

I may be here - map location.

QR of my Contact Details:

**Other Professional Activities**

You might be looking for some web sites I am responsible for:

- RKB Explorer
- sameAs.org
- See UK

**Hugh works with:**

- Dr Ian Millard (explain)
- Professor Nigel R Shadbolt (explain)
- Afraz O Jaffri (explain)
- Dr Nicholas Gibbins (explain)
- Bene Rodriguez-Castro (explain)
- Professor David C De Roure (explain)
- Dr Les A Carr (explain)
- dr monica mc schraefel (explain)
- Dr Harith Alani (explain)
- Dr David E Millard (explain)
- Dr. Yvonne Howard (explain)
- Dr. Manuel Salvadores (explain)
- Tope Omitola (explain)
- Gianluca Correndo (explain)

This list of people is sourced from RKBExplorer, the result of an ECS project. Click on the "explain" link to the right of each person to find out how they are connected.
**Note**: the explain link requires heavy processing and may take some time to load.

Tel: +44 (0)23 8059 6000 | Contact ECS | Intranet | © University of Southampton

*Figure 1: Example of researcher profile from the University of Southampton*

At present there is little coherence amongst the academic community in how a researcher presents personal data online. In general a webpage, hosted by the researcher's institution, is the most prevalent solution but these generally maintain little consistency, even between departments. There are a multitude of other solutions ranging from ungoverned social networks such as LinkedIn [2] to more dedicated systems such as PIMS [3].

## Community Stakeholders

This report identifies four key groups of stakeholders within the academic community that are currently driving demand for researcher profiles. These are:

### Funders and Government agencies

Those organisations funding or providing services to academic research in the UK have a vested interest in tracking researchers. Funders - including research councils, government bodies, and some Non-profit organisations - benefit from a researcher profile system through more efficient assignment of funds and more accurate knowledge about the recipients of their finance. Demand from funding organisations is driven by the need for greater granularity of information and avoidance of double spending on projects.

Government agencies tasked with providing HE services nationally also require detailed information about researchers to provide many of their services. Of primary importance is national reporting functionality - for example gathering data to output as reports for the REF or HESA reports. This data is also seen as vital in creating more accurate roadmaps and strategic planning.

**HEIs**
The institutions - mainly universities and colleges - that employ researchers need to identify individuals accurately. There is already a good deal of infrastructure in place to identify employees for payroll or accounting purposes but more advanced researcher profiles offer additional benefits. An advantage to institutions is in support for internal reporting requirements such as Staff Performance Review. Institutions may also benefit from increased exposure and publicity generated by researcher's profiles. HEIs face issues when researchers move between institutions or work for several different organisations at the same time.

**Commercial interests**
Commercial companies involved in the academic sector are showing significant demand for researcher profiles driven by cost savings and other efficiencies that a unified profile system could present. Commercial publishers in particular stand to gain significant revenue upside if they are successful in tying researcher profile information into proprietary publishing systems. An example of this is the Thomson-Reuters Researcher ID system which is linked to the Web of Science commercial portal.

**Researchers**
In general the main demand from researchers is in promoting their research and gaining reputation in the community. In defining demand from this group however there are necessarily a wide range of drivers as the needs of researchers vary depending on who and where they are. Factors include: faculty, research area, stage in career, geographical location.

# Overview of data sources

A data source will normally be managed by an organisation that controls what data appears, the format and the frequency which data is updated. The organisation has an influential role in deciding how a researcher is represented and what information is made available.

Data sources can be categorised based on the scope of the organisation managing the data source. This can be viewed as a spectrum ranging from fully global systems where the decision processes for change are based on high level strategy down to institutional systems which service local needs. These categorisations should be viewed as indicative only.

## Host vs Producer Considerations

It is worth noting that there is a distinction between the organisation which controls data usability and visibility, and an individual or organisation that actually produces and provides the data to that system. For example, publishers host and disseminate journal content, but are reliant on researchers to produce papers for them to publish.

Another example is that of social networks, which are run by global commercial organisations, but all of the content is provided by individuals. Although the managing organisation is responsible for hosting the information, they will provide only basic verification and the reliability of information is therefore dependent on the individual controlling the profile.

## Commercial vs Public Considerations

While there is no clear definition of what constitutes a public system (with issues complicated by national boundaries and private/public partnerships) - it can be agreed that commercial systems are distinct from not-for-profit. Commercial systems generally have implications in terms of proprietary aspects of the systems.

The difference between commercial and not-for-profit systems mainly manifests itself in terms of data licensing models. Most commercial systems will aim to charge subscriptions for access to data, making those data sources limited to a select group of users. Besides control of data, the licensing of the software code used to build a system is also a potential issue. From a technical perspective this can create issues with interoperability and usability of systems in different parts of the sector.

In order to be successful any identifier system must engage both commercial and public organisations. It is highly unlikely that any system can attain global reach and relevancy without buy-in from all forms of publishers and content creators.

## Global Services

At the international level the majority of data sources are focused on publication and article metadata. Some of these may be global in reach but only focused on a subset of scholarly output. Nearly all commercial gatekeepers tend to be at this level due to the global nature of information provision.

| Service | Description | Data |
|---|---|---|
| AuthorClaim [4] | AuthorClaim links scholars with records about works they have written | Person identifiers, bibliographic metadata |
| Crossref [5] | Citation linking backbone for all scholarly information in electronic form | Article identifiers (DOI), and some bibliographic metadata |
| Google Scholar Citation profiles [6] | Commercial author profile system operated by Google (still at an early development stage) | Bibliographic metadata, person identifiers, article identifiers |
| Incites [7] | Thomson-Reuters Citation Data and other metrics | bibliographic metadata, citation data |
| Mendeley [8] | Personal reference management service | bibliographic metadata |
| Microsoft Academic Author Profile [9] | Commercial author profile system operated by Microsoft (still at early development stage) | Person identifiers, article identifiers, bibliographic metadata |
| ORCID [10] | Global researcher identifier system (proposed) | (TBC) Person identifiers, article identifiers and researcher profiles |
| Pubmed [11] | Free database accessing MEDLINE database of references and abstracts on life sciences and biomedical topics | Person identifiers, article identifiers, bibliographic metadata |
| Social Networks | Social profiles hosted by various commercial and not-for-profit organizations | Various |
| Scopus Author | Commercial author identifer system operated | Person identifiers, article |

| | | |
|---|---|---|
| Identifier [12] | by Elsevier B.V | identifiers, bibliographic metadata |
| Subject Repositories | Subject focused repositories storing data relevant to particular research fields. e.g arXiv [13], RePEC [14] | Person identifiers, article identifiers, bibliographic metadata, full-text |
| VIAF [15] | Matches and links widely-used library authority files | Library authority files |
| Web of Knowledge [16] | Academic citation indexing and search service | Person identifiers, article identifiers, bibliographic metadata |

*Table 1: International level datasources*

## National services

There are a number of national level efforts to provide metadata about researchers in particular countries. These vary in scope and functionality and as a result reliability of data from these sources varies. They are normally determined national funding sturcutres.

| Service | Description | Data |
|---|---|---|
| British Library Auth data [17] | Metadata relating to British Library records | Bibliographic metadata |
| Digital Author Identifier [18] | National identifier system - Netherlands | Person identifiers |
| DissOnline [19] | Theses and Dissertations service - Germany | Thesis metadata |
| CRISTIN [20] | National research information system - Norway | Person identifiers, bibliographic data, full-text |
| HESA | National statistics agency metadata | Person identifiers, funding metadata |
| LATTES [21] | Curricula, research groups and institutional database - Brazil | Organisational data, and potentially person identifiers |
| People Australia [22] | Author Identifier system - Australia | Person identifiers, bibliographic metadata |
| Researcher Name Resolver [23] | Name authority service - Japan | Person identifiers |
| Names project [24] | UK based name authority service lead by MIMAS | Person identifiers |
| PIMS | Grant/ Funding Tracking system developed by JISC | Person identifiers, Project Identifiers, Funding metadata |
| Research councils [25] | Research Councils such as EPSRC [26] or BBSRC [27] hold metadata about | Person identifiers, Project Identifiers, Funding metadata |

| | | |
|---|---|---|
| | researchers working on projects | |
| Wellcome trust [28] | Funding body holding metadata about researhcers involved in funded projects | Person identifiers, funding metadata |

*Table 2: National services*

## Institutional services and systems

At the institutional level there are a wide variety of systems developed to meet individual needs. These will often duplicate effort or contain identifiers and data that are not easily accessible by outside users.

| Service | Description | Data |
|---|---|---|
| Finance/HR ids | metadata held by administrative departments for authentication and tracking | Person identifiers |
| Institutional CRIS | Current Research Information System. Examples include Symplectic [29] and Atira [30] and Avedas [31] | Person identifiers, Bibliographic data, organisational data, funding metadata |
| Institutional repositories | Silos for institutional publications | Bibliographic data, full-text articles |
| Institutional webspages | Institutions may host individual profiles for staff, relevant metadata or linked open data relating to researchers | Person identifiers, Bibliographic data, organisational data, funding metadata |
| Library user lists | metadata held by libraries for authentication and tracking | Person identifiers |
| Project specific pages | Webpages describing research projects across several instiutions | Person identifiers, funding metadata, project metadata |
| VIVO [32] | open source, semantic web application originally developed and implemented at Cornell | Person identifiers, person metadata, bibliographic metadata, funding metadata |

*Table 3: Institutional services and systems*

## Auth NZ systems

| Service | Description | Data |
|---|---|---|
| Athens [33] | Access and Identity Management service | Authentication credentials, possibly useful as person identifiers |

| | | |
|---|---|---|
| OpenID [34] | Provides mechanism to simplify multiple log-in and identify users | Authentication credentials, possibly useful as person identifiers |
| Shibboleth [35] | Open source service for web single sign-on across organisational boundaries | Authentication credentials, possibly useful as person identifiers |

*Table 4: Institutional data sources*

# Case studies

This section outlines eight case studies pertinent to discussion of researcher identifiers and profiles. They have been chosen to highlight key aspects and choices presented by current systems, as well as the different kinds of data available, rather than to provide an exhaustive list.

The case studies aim to delineate what is available from current data sources as well as some common barriers to use. They include a brief description of the system, data held, technology model and reasons to use/ avoid the system.

One of the key dimensions to consider is which organisation or body acts as gatekeeper for the data. The sustainability of any system is tied to issues of control and gatekeeping and this is ultimately tied to funding. In many cases the funding body and gatekeeper will be one and the same however it may be more complicated with multiple organisations responsible for various parts of the workflow.

| **ORCID** |
|---|
| **Description**<br>The ORCID Initiative combines public and commercial partners to address multiple needs for researcher identifiers. It is a joint effort aimed at creating an open, independent registry to be used as an industry standard. Its main focus will be on resolving ambiguities by assigning unique identifiers linkable to a researcher profile. This will improve discovery of research and the efficiency of funding and collaboration. It is currently under development and is projected to be launched in its initial phase in 2012. |
| **Holder organisation**<br>ORCID is managed by a board of directors from 14 organisations representing both public and private concerns. These are:<br>● Association for Computing Machinery (ACM)<br>● Cornell University<br>● CERN<br>● CrossRef<br>● Elsevier<br>● Hannover Medical School<br>● Harvard University<br>● Online Computer Library Center (OCLC)<br>● National Institute of Informatics (NII)<br>● Nature Publishing Group (NPG)<br>● MIT Libraries<br>● Thomson Reuters<br>● Wellcome Trust<br>● Wiley-Blackwell |
| **Data held**<br>The system is still under-development however the aims are wide ranging with a view to include all relevant metadata.<br>● Researcher Metadata - Name, DoB, institution, institution, city, etc.<br>● Article metadata - title, journal publications, citations, bibliographic data,<br>● Author Identifier - ORCIDs - unique reseacher identifier string |
| **Technology Model** |

Currently in development - The project has a stated aim to use Open Source code throughout and give researchers access to modify personal data. Phase 1 development is based on the codebase from Thomson-Reuters Researcher ID service.

**Reasons to use this system**
Although ORCID is still in the development phase, proposed uses for the system include:

- Researchers can present their research output publicly and generate interest in their work
- A list of a given researcher's publications can be generated to understand their career output
- Grants and awards received by researcher or group can be searched
- Research output from a given institution can be generated to analyse strengths and weakneses
- Papers published as a result of specific funding can be viewed
- Authors and reviewers can be tracked more accurately in journal submission systems

**Business Model - Who pays for it?**
Once launched, the system will be funded by a range of fees although the exact cost structure is still not finalised. Charges are envisioned to include: Membership fees, transaction fees, fee-for-service (real-time access, alerts, bulk querying, disambiguation)

Funding to date has come from a range of sources with the main costs shouldered by founder organisation. While still pre-launch ORCID has been funded by two sponsorship rounds, grants and in kind work.

In kind (2010) – staff time; board, working groups, technical and legal
Sponsorship 1 (early 2011) - $244,000 in donations from Founding Sponsors
Small grants (mid-2011) – Mellon ($49,000); VIVO ($25,000)
Sponsorship 2 (late 2011) – announced in August to get additional $250,000

**Barriers to implementation as a data source**
The stated business plan includes a notion that organisations would pay for an upper tier of access to the data. It is unclear until the service launches what this would mean for an aggregating service that seeks to republish this data under different financial terms, including providing a similar level of service for free.

At worst, a service would be free to use the annual planned release of ORCID profile information as this would be under a CC0. Until the terms and conditions of higher tier access is clarified, it is impossible to comment further.

## Scopus Author Identifier

**Description**

The Scopus Author Identifier aims to disambiguate author references. It is a commercial system funded by subscription and paid for content. It aims to distinguish between articles belonging to authors with similar names, increase confidence that results cover all output across variations of a single name, and account for different representations in foreign languages.

**Holder organisation**

Scopus is operated by Sciverse [36]. SciVerse is a registered trademark of Elsevier Properties S.A [37]

**Data held**

- Researcher Metadata - Name, institution, subject area etc.
- Article metadata - title, journal publications, bibliographic data
- Author identifier - Scopus author ID

**Technology Model**

Scopus is based on proprietary technology implemented by Elsevier to support its Sciverse online retail division.

**Reasons to use this system**

- Researchers can present their research output publicly and generate interest in their work
- A list of a researcher’s publications can be generated to understand their career output
- Authors and reviewers can be tracked more accurately in Sciverse submission system

**Business Model - Who pays for it?**

Elsevier operates an extensive retail business delivering content and technology for the scientific sector via its Sciverse and related online portals. Scopus Author Identifier is funded from the profits of ancillary operations.

**Barriers to implementation as a data source**

Elsevier sells access to the Scopus database and as such, they may be amenable to embedding search results or individual author profile information widgets or pages given a suitable subscription but building on, repurposing and republishing this information might be difficult to negotiate.

## PubMed Central

**Description**

PubMed Central is a free digital database of full-text scientific literature in biomedical and life sciences. It predominantly draws from MEDLINE, life science journals, and online books [38]. It has grown from the Entrez PubMed biomedical literature search system. The system applies a unique PubMed identifier (PMID) to each PubMed record.

**Holder organisation**

PubMed Central is maintained by PMC International (PMCI) - a collaborative effort between NIH [39] and NLM [40], the publishers whose journal content makes up the PMC archive, and organizations in other countries that share NIH's and NLM's interest in archiving life sciences literature. In the UK, UK PubMed Central (UKPMC) [41], is funded by the NHS [42] and Medical Research Council [43] amongst other research funders. It relies on information from commercial medical publishers but is majority publicly funded.

**Data held**

- Researcher Metadata - Author Name, institution,
- Medical Subject Headings (MeSH) metadata
- Article metadata - title, journal publications, citations, bibliographic data,
- Article Identifier - PMIDs - unique identifier string

**Technology Model**

Articles are sent to PubMed Central in XML or SGML, with many publishers using the NLM Journal Publishing DTD [44]. Received articles are converted via XSLT to the NLM Archiving and Interchange DTD. Graphics are converted to standard formats and sizes. Bibliographic citations are parsed and automatically linked to the relevant abstracts.

**Reasons to use this system**

- It is compulsory for NIH funded research
- A list of a researcher’s publications can be generated to understand their career output
- Authors and reviewers can be tracked more accurately in journal submission systems

**Business Model - Who pays for it?**

Pubmed is fully funded by the NIH as stipulated by US law that requires that all research funded by NIH is made available Open Access via the system. It also relies on some data and support from commercial publishers focused on author-pays business models.

**Barriers to implementation as a data source**

No author disambiguation or resolution is yet done and whilst much of the core metadata is available openly, this work would still have to be carried out before it could be used as a source to augment author profiles without a user selecting which works they authored.

## JISC Programme Information Management System (PIMS)

**Description**
The Programme Information Management System (PIMS) is a record of all JISC-funded programmes and projects, from the year 2000 onwards.

**Holder organisation**
JISC

**Data held**

- Researcher Metadata - Author Name, institution,
- Project metadata - Funding strand, progress state, etc

**Technology model**
PIMS is based on bespoke code maintained by JISC.

**Reasons to use this system**

- Papers published as a result of specific funding can be viewed
- Grants and awards received by researchers or group can be searched

**Business Model - Who pays for it?**
PIMS is funded internally by JISC and is made freely available as a public resource. JISC is a national body funded by UK higher and further education funding bodies and research councils through an annual budget recommendation process.

**Barriers to implementation as a data source**
There is no formal commitment to maintain the integrity of the personal identifiers used for the primary investigators; no explicit responsibility for the uniqueness of the identifier, nor that the identifier will not change in time. These may be critical concerns for the service, but requires a clear statement to this effect.

## AuthorClaim

**Description**
The AuthorClaim registration service is a self-claim identifier system operated on a not-for-profit basis. It allows researchers to identify themselves with work held in the system's bibliographic database. It states its aims as creating and promoting profiles for researchers, disambiguating researchers with similar names, and providing statistics to enable rankings.

**Holder organisation**
Thomas Krichel - Independent developer and researcher in digital libraries. He is the author of the RePEC digital library for economics

**Data held**

- Researcher Metadata - Author Name, institution,
- Article metadata - title, journal publications,

**Technology Model**
AuthorClaim is a clone of the RePEc Author Service with a major focus on researchers in economics. It is based on a proprietary database of research outputs against which authors can claim.

**Reasons to use this system**

- Researchers can present their research output publicly and generate interest in their work
- A list of a researcher's publications can be generated to understand their career output

**Business Model - Who pays for it?**
AuthorClaim is run by Thomas Krichel as a not-for-profit tool. The development of the software for the service was funded by an Open Society Institute grant to the ACIS project.

**Barriers to implementation as a data source**
Material cannot be submitted directly to AuthorClaim. All editorial decisions are controlled by one individual at the moment, but this is potentially due to a shortage of resources alone.

## University of Southampton ECS Profile Pages

**Description**
The University of Southampton provides some staff with a configurable profile page and unique URL [45]. A good example of this is the Electronics and Computer Science department. Features and presentation vary between departments.

**Holder organisation**
University of Southampton

**Data held**

- Researcher metadata - Name, contact details, biographic information, previous publications
- Article metadata - citations, bibliographic data, colleagues

**Technology Model**
Staff member’s profiles are based mainly on static HTML updated on an ad-hoc basis by researchers themselves. There are some dynamic elements but these are not integral. The collaborators list, for example, is sourced from RKBExplorer [46].

**Reasons to use this system**

- Researchers can present their research output publicly and generate interest in their work
- A list of a researcher’s publications can be generated to understand their career output
- It provides a mechanism to find collaborators that a researcher works with
- Discover projects and seminars that a researcher is involved with

**Business Model - Who pays for it?**
Recently the University decided that this profile system was outside of the remit of the services it should provide and has ceased to provide support. Although profiles are still available there will necessarily be degradation of service going forward. Staff member profiles were hosted by the institution. Researcher’s updated content by contacting system administrators. Hosting and staff costs were absorbed by the institution.

**Barriers to implementation as a data source**
Although the decision to discontinue service is unfortunate it serves as an example of the risk inherent to institutionally supported identification services, regardless to their technological merit or usefulness to researchers.

## arXiv

**Description**
arXiv is a repository providing data and articles for a number of scientific research fields. Submissions to arXiv must conform to Cornell University academic standards.

**Holder organisation**
Cornell University - a private not-for-profit educational institution

**Data held**

- Researcher metadata - Name, contact details, biographic information, administrative metadata
- Article metadata - title, full text, journal publications, citations, bibliographic data,

**Technology Model**
arXiv is is a highly-automated electronic archive and distribution server for research articles. It has previously been based on a bespoke technology, but is currently migrating to the Invenio [47] platform from CERN.

**Reasons to use this system**

- A list of a researcher's publications can be generated to understand career output
- Research output from a given institution can be generated to analyze strengths and weaknesses
- Authors and reviewers can be tracked more accurately in journal submission systems

**Business Model - Who pays for it?**
arXiv is funded by Cornell University Library and by supporting user institutions. The National Science Foundation funds research and development by Cornell Information Science.

**Barriers to implentation as a data source**
Organisations may be wary of submitting to much control over metadata to an instituion that may be seen as a competitor.

## HESA - Staff Individualised Record

**Description**
The Higher Education Statistics Agency (HESA) gathers various identifier information on staff employed by academic institutions in the UK as part of its remit to generate relevant statistics.

**Holder organisation**
Higher Education Statistics Agency (HESA)

**Data held**

- Researcher metadata - Name, contact details, biographic information, administrative metadata

**Technology model**
HESA uses a bespoke solution to track researcher metadata. Collaborating universities and institutions append unique identifier strings to data that they submit to HESA. Details on how to generate identifier strings are available on HESA's website but researcher identifier information once submitted is not shared by HESA.

**Reasons to use this system**
This system is compulsary for certain institutions that are publicly funded in the UK. HESA is the official agency for the collection, analysis and dissemination of quantitative information about higher education in the UK. This is primarily statistics related to staff numbers and finance in the UK HE sector.

**Business Model - Who pays for it?**
HESA is a private limited company which has formal agreements with government departments to provide the data which they require. It is funded by subscription from all of the universities and higher education colleges throughout the United Kingdom. Publications are available for sale via HESA's website.

**Barriers to implentation as a data source**
An Institution's agreement with HESA may only require limited data sets to be submitted. Not all researchers may be classified within the HESA remit.

# Issues Identified

**Gatekeeping and community stakeholders**
A key consideration of any data source is who or what acts as gatekeeper for the data. This may be more complex than a single organisation and will invariably be linked to funding mechanisms. It is vital that any system has sufficient buy-in from all relevant parties to ensure there is drive to populate the system. Any system that does not consider the aims of all four community stakeholder groups is high risk as excluded parties will ultimately have no motivation to work with that system.

Equally, if there is input from too many disparate organisations there will be an impediment to success as differing opinions will delay progress and uptake. If control of data sources is devolved too far there will be significant issues in terms of interoperability, acceptance and uptake.

**Sustainability**
The continued functioning of any system rests on two factors - financing and buy-in. If any system generates running costs that exceed available budget then it will not continue. Similarly if any system becomes irrelevant to the interests of any party then that party will cease to use it.

Ensuring continuity is a key concern. This is highlighted by the example of The University of Southampton’s profile pages which, although fit for purpose, will no longer be supported by the university. This decision reflects the fact that the cost of hosting such pages does not explicitly support the university's mission. It is therefore important that delegation of control for any data source is given to a suitably purposed organisation capable of raising sufficient finance to meet its costs.

At present the majority of systems considered relevant to this report are based around Western academic research institutions and publications. As a result the majority of stakeholders and gatekeepers mentioned are either national or commercial organisations located in Western countries. However, there is undoubtedly significant need for similar functionality in other economies, and how this need is met may impact sustainability - either in terms of increased cost or reduced global buy-in.

For systems that are required to identify researchers globally the geographical mix will likely change. As such, some thought to account for changing funding and service requirements is necessary; otherwise Western systems face becoming irrelevant, or simply replaced, if they do not meet the needs of a global audience.

**Walled gardens**
Commercial subscription systems face issues ensuring parity between sectors, institutions, and geographies. There is a risk that the scope of representation of researchers identified by a closed commercial system may not fully represent the true breadth of research activities, and may even serve to cast undue doubt on the respectability or validity of unidentified research. Taking the term ‘researcher’ in the broader sense there are many professionals that may exist outside the definitions created by commercial organisations as they do not generate revenue for the publisher.

**Privacy**
A major concern from the researcher perspective is to ensure that their data are not given to people without their consent. This could be personal data that they do not want publicly available in any form or more nuanced data that they wish to withhold for particular reasons.

Researchers may wish to keep some data about their research area private if it is senstive or if they feel it may give away a competitive advantage. For example, researchers involved with animal testing may request that their research projects are private to avoid exposure to animal rights activists.

This is a potentially significant legal issue for those organisations managing the system. Should a researcher feel that private data was published without consent there is scope for legal action against the holder organisation.

**Technology**
The technology employed by a data source has a direct impact on usability and functionality. The lack of an API or well recognised standards can impact the interoperability of any service, and while data may in theory be available it is only useful if it can be retrieved in a workable format.

**Disambiguation**
While individual systems may provide disambiguation services they will not always guarantee that it is possible to disambiguate across multiple data sources, as data quality and disambiguation data may vary.

# Recommendations

There are a number of data sources that demonstrate best practice, and building on these can provide quick wins toward a researcher profile systems infrastructure.

## Best practice recommendations

- PubMed is identified as a data source that exemplifies best practice in terms of openness and sustainability.
- HESA exemplifies a successful top down approach where institutions are required to supply data at the UK level. This is an example of a data source with good reach and buy-in.
- Crossref is identified as a good example of a gatekeeper organization, bringing in cross-party buy-in and engagement and helping to raise sustainable financing.
- Large subject repositories, such as RePEC or arXiv, are identified as strong data sources for publications information and provide current best practice for openness in scholarly output.

## Community engagement recommendations

It is likely that individual HEIs and researchers will not have the resource to commit to a long term data sources strategy, but will tend to engage with particular data sources on an as needed basis. So in order to ensure that information held within a particular data source is well maintained and kept up to date, specific effort must be made by overarching organisations such as JISC to ensure long term stakeholder engagement. These efforts should include:

- Encourage and support open discussion and consultation
- Offer the opportunity for representatives of all stakeholder groups to provide input
- Maintain dialogue with HEIs and provide incentive for researchers to contribute

## Next steps recommendations

- Before committing to any particular data source as a provider of national researcher profile information, a full technical and policy review of a shortlist of data sources should be performed.
- ORCID should be carefully considered for possible adoption. It is an as yet incomplete technology and framework, but closer comparison of use cases may show that it may in principle be a close match for the needs of the UK HE sector.
- Wide engagement on an effort to identify key use case scenarios should inform these

decisions.

## Concluding remarks

As particularly demonstrated by the long list of already available data sources included in this report, there is no shortage of information, or of sources from which to retrieve it; and as the accompanying technical report describes, many solutions are available for supporting the dissemination of this data further. The critical factor in meeting the requirements of the HE sector will be in carefully identifying the critical use cases that people and institutions really need, in demonstrating why meeting those use cases will make our lives better, and then implementing the solution that meets those use cases. This will involve not only technical implementation, but implementation of changes to the social infrastructures by which we achieve our aims.

# References

1. HESA - UK - http://www.hesa.ac.uk
2. LinkedIn - http://www.linkedin.com/
3. JISC - PIMS - https://pims.jisc.ac.uk/
4. AuthorClaim- http://authorclaim.org/
5. Crossref - http://www.crossref.org/
6. Google scholar - http://scholar.google.co.uk/
7. Incites - Thomson-Reuters Citation Data - http://researchanalytics.thomsonreuters.com/incites/
8. Mendeley data -http://www.mendeley.com/
9. Microsoft Academic search - http://academic.research.microsoft.com/
10. ORCID - http://orcid.org/
11. Pubmed Central- http://www.ncbi.nlm.nih.gov/pmc/
12. Scopus Author Identifier - http://www.info.sciverse.com/scopus/scopus-in-detail/tools/authoridentifier/
13. arXiv - http://arxiv.org/
14. RePEC - http://repec.org/
15. VIAF -http://viaf.org/
16. Web of Knowledge - http://apps.isiknowledge.com/
17. British Library Auth data -http://www.bl.uk/bibliographic/data.html
18. Digital Author Identifier - SURFfoundation - Netherlands - http://www.surffoundation.nl/en/themas/openonderzoek/infrastructuur/Pages/digitalauthoridentifierdai.aspx
19. DissOnline - Germany -http://www.dissonline.de/
20. CRISTIN - Norway - http://www.cristin.no/as/WebObjects/cristin
21. LATTES- Brazil - http://lattes.cnpq.br/
22. People Australia - Australia - http://www.nla.gov.au/initiatives/peopleaustralia/
23. Researcher Name Resolver - Japan - http://rns.nii.ac.jp/
24. Names project - http://names.mimas.ac.uk/
25. Research councils - http://www.rcuk.ac.uk/Pages/Home.aspx
26. EPSRC http://www.epsrc.ac.uk
27. BBSRC - http://www.bbsrc.ac.uk/
28. Wellcome trust - http://www.wellcome.ac.uk/
29. Symplectic - http://www.symplectic.co.uk/
30. Atira -http://atira.dk
31. Avedas - http://www.avedas.com/
32. VIVO - http://vivo.slis.indiana.edu/
33. Athens - https://auth.athensams.net/
34. OpenID - http://openid.net/
35. Shibboleth/IdP - http://shibboleth.internet2.edu/
36. Sciverse - http://www.hub.sciverse.com/action/home/proceed
37. Elsevier - http://www.elsevier.com/
38. Pubmed Central resources - http://www.nlm.nih.gov/bsd/pmresources.html
39. National Institutes of Health - http://www.nih.gov/
40. National Library of Medicine - http://www.nlm.nih.gov/
41. UK PubMed Central (UKPMC) - http://ukpmc.ac.uk/
42. NHS - http://www.nhs.uk/
43. Medical Research Council - http://www.mrc.ac.uk
44. NLM Journal Publishing DTD - http://dtd.nlm.nih.gov/publishing/
45. Southampton researcher profile system - http://www.ecs.soton.ac.uk/people/
46. RKB Explorer - http://www.rkbexplorer.com/
47. Invenio - http://invenio-software.org/