INTERNATIONAL STANDARD NAME IDENTIFIER (ISNI)

THE OPEN RESEARCHER & CONTRIBUTOR ID (ORCID)

THE NAMES PROJECT

STANDARD ADDRESS NUMBER (SAN)

INSTITUTIONAL IDENTIFIERS (I²) AND ISNI
Researchers and library patrons are increasingly expecting instant access to the information they need. While the availability of electronic content grows daily and standards such as OpenURL have drastically improved discovery, impediments still remain. At NISO, a number of current projects are underway to improve discovery, access, and delivery of content:

**KBART: Phase II**

**PHASE I** of the joint NISO/UKSG KBART (Knowledge Bases and Related Tools) project resulted in practical recommendations for exchanging metadata between content providers and knowledge base developers. These recommendations are intuitive, easy for content providers to implement, and easy for knowledge base developers to process.

**PHASE II** builds on that work to focus on the more advanced, complex issues that cause problems in this area. Learn how to implement the recommendations from Phase I and about the next stage of this work at [www.niso.org/workrooms/kbart](http://www.niso.org/workrooms/kbart).

**IOTA: Improving OpenURLs Through Analytics**

IOTA is a two-year project to investigate the feasibility of creating industry-wide, transparent, and scalable metrics for evaluating and comparing the quality of OpenURL implementations across content providers. At this time, nearly 9 million OpenURLs have been analyzed from log files. The reports created from this analysis allow publishers to see where they can make improvements to their OpenURL strings so that the maximum number of OpenURL requests can be resolved—bringing more readers to their products. Visit [openurlquality.niso.org](http://openurlquality.niso.org) to view the metrics and learn how to add your data to the project. Find out more at [www.niso.org/workrooms/openurlquality](http://www.niso.org/workrooms/openurlquality).

**Presentation and Identification of E-Journals (PIE-J)**

Unless journal websites accurately and uniformly list all the titles under which content was published, user access to desired journal articles is considerably diminished. When journals change titles or publishers, their content must remain easily accessible. This new working group will be developing recommendations that will provide much-needed guidance on the presentation of e-journals to publishers and platform providers—particularly in the areas of title presentation, accurate use of the ISSN, and citation practices—that will solve some long-standing concerns of serials librarians. See [www.niso.org/workrooms/ejournalpresentation/](http://www.niso.org/workrooms/ejournalpresentation/) for more information.

**ESPReSSO: Establishing Suggested Practices Regarding Single Sign-On**

This NISO Chair’s Initiative was launched to develop recommendations that will improve the user experience when using diverse electronic services by providing transparent single sign-on authentication across distributed service providers. The end result of this work will be small, smart conventions for moving the user within a session from one licensed site to another, so that publisher content can be accessed easily and seamlessly. Find out more at [www.niso.org/workrooms/sso](http://www.niso.org/workrooms/sso).

**DO YOU HAVE A SUGGESTION FOR NEW WORK? WE’D LIKE TO HEAR FROM YOU!**

[www.niso.org/standards/suggest](http://www.niso.org/standards/suggest) OR visit [www.niso.org/workrooms](http://www.niso.org/workrooms) for more information.
InformatIon StandardS Quarterly (ISQ) is a publication by the national Information Standards organization (nISo). ISQ is nISo’s print and electronic magazine for communicating standards-based technology and best practices in library, publishing, and information technology, particularly where these three areas overlap. ISQ reports on the progress of active developments and also on implementations, case studies, and best practices that show potentially replicable efforts.

CONTENTS

3 From Guest Content Editor

FE 4
FEATURE
4 ISNI: A New System for Name Identification
10 ORCID: Unique Identifiers for Authors and Contributors

IP 14
IN PRACTICE
14 The Names Project: A New Approach to Name Authority

OP 20
OPINION
20 Identify This! Identifiers and Trust

SP 23
SPOTLIGHT
23 The Use of the Standard Address Number (SAN) in the Supply Chain

NR 26
NISO REPORTS
26 IP and ISNI: Improving the Information Supply Chain with Standard Institutional Identifiers

NW 30
NOTEWORTHY

SD 33
STANDARDS IN DEVELOPMENT
SEPTEMBER
14  Discoverable, Accessible, Available: Preserving Digital Content (NISO Webinar)
28  ROI in Linking the Semantic Web (NISO Webinar)

OCTOBER
NISO Two-Part Webinar: Managing Data for Scholarly Communications
  12  Part 1: Supplemental Materials
  19  Part 2: Technical Management
24–25  The E-book Renaissance (NISO Two-Day Forum Baltimore, MD)

NOVEMBER
9   New Discovery Tools (NISO Webinar)
16   The RDA Vocabularies: Implementation, Extension, and Mapping (NISO/DCMI Webinar)

DECEMBER
14  Assessment Metrics (NISO Webinar)

NISO OPEN TELECONFERENCES
Join us each month for NISO’s Open Teleconferences—an ongoing series of calls held on the second Monday of each month as a way to keep the community informed of NISO’s activities. The calls also provide an opportunity for you to give feedback to NISO on our activities or make suggestions about new activities we should be engaging in.

The call is free and anyone is welcome to participate in the conversation. All calls are held from 3:00 – 4:00 p.m. Eastern time.

Webinar Subscription Package Discounts
Buy all 14 for the price of 7!
Buy 4 and get 3 free.
Identifiers have always been essential standards in libraries and publishing. Content identifiers, such as the ISBN, are well established, but institutional and personal identifiers are much newer, the need having emerged as more and more content is available electronically. This issue of ISQ looks at existing and new institutional and people identifiers, how they have developed, and what new initiatives are underway in this area.

My interest in identifiers started in my initial training as a geologist and my subsequent library and information experience over 40 years. I have always wanted to put things into categories or “boxes.” In 1996, I was working with a team developing an innovative hosting platform and had lively discussions with many people about how the arrival of scholarly electronic publication meant that it was time to start systematically identifying users and usage. I suggested that if you could have identifiers for the content, the user, the rights, and the time period, you could automatically provide access to everything one was entitled to see. Every January the listservs demonstrate that this still isn’t the case, with plaintive pleas asking who has lost access to what.

However, 15 years later it seems that the time may have finally come when there is a universal recognition of this requirement. GEOFF BILDER expresses this eloquently in his opinion piece pointing out the importance of trust in the electronic world. I was involved in the initial Journal Supply Chain pilot which tried to highlight the “points of pain” in the process of getting e-journal content from the supplier to the user. This pilot was picked up by NISO and evolved into the Institutional Identifiers (I) Working Group. JODY DE RIDDER summarizes the findings of this Working Group and their final selection of the International Standard Name Identifier (ISNI) as their working identifier. The development of the ISNI has been a thorough and comprehensive exercise within the International Organization for Standardization (ISO), with many parties having interests and needing their requirements met. JANIFER GATENBY of OCLC has been a key part of that initiative and has also contributed considerably to the I Working Group. She and ANDREW MCEWAN of the British Library describe the architecture and operation of the ISNI system, which will bring together many different name identification registries. LOUISE TIMKO makes a plea for the “less glamorous” but no less essential SAN (Standard Address Number), which is a key part of the publishing supply chain.

Identification of researchers has been a hot topic for some years and with bigger investments in institutional repositories this need has only increased. The JISC funded Names Project is summarized by ALAN DASKIN, AMANDA HILL, and DANIEL NEEDHAM. Publishers have started many fragmented author identifier schemes, but finally there is some collaboration between them and the academic community. MARTIN FENNER describes ORCID (The Open Researcher and Contributor ID), the ambitious initiative to bring all these schemes together and perhaps even work with the ISNI.

The papers in this issue do show that not only has the need for institutional and personal identifiers been recognized, but there is increasing collaboration between the constituent parts of the scholarly community to make sure that not only are the standards developed but they are also implemented.

The company I founded, Ringgold, has a continuing interest in institutional identification, but I am now moving back to my consultancy roots where I can put things into boxes for others. doi: 10.3789/isqv23n3.2011.01

Helen Henderson | Managing Director at Information Power Ltd
If I like a novel of one author, I look to see what else he or she has written. The same is true if I find somebody who writes well on topics that interest me, so I would like to find all he or she has written in any form—whether book, article, or blog. The same applies to all creators—composers, performers, film directors, researchers, artists, producers, publishers, and others.

The importance of the creator in determining selection is clearly important to me; titles give me an idea about subject whereas creators give me an idea about quality and pertinence. Yet existing online databases including search engines, online book shops, and to a lesser extent library catalogs, are poor in grouping by creator. In most Internet sites, when you click on “more by this author, composer, etc.” typically a keyword search is launched that returns both imprecise and incomplete results. The user cannot be sure that what is proposed is, in fact, created by the same author. Further searches are often necessary for identity identification.

Identification of creators and other contributors is also critical to societies who administer rights information and royalty payments. These societies in their day-to-day work to channel the incoming royalty payments, often find themselves in a position of searching for information to accurately identify the rights holders. As the market is shifting from physical towards digital...
and from purchased copy to licensed access, the accurate and unambiguous identification of the parties involved is even more important, not only to rights management systems but also to trade organizations, publishers, aggregators, distributors, and retailers. The current efforts of libraries, rights management societies, and trade organizations to disambiguate creators are arduous, and their efforts are duplicated.

Identifiers serve as shorthand for the metadata that differentiate one identity from another. The idea of a simple unique identifier that everyone in the world could use to identify the same person, corporate body, or similar entity has been proposed many times during the last 30 years. Finally, an international standard (ISO 27729) for an International Standard Name Identifier (ISNI) has been approved to respond to this need. ISNI has been designed as a bridge between existing proprietary right holder identification systems, such as the Interested Party Identifier, and resource discovery tools, such as the Virtual International Authority File (VIAF). By sharing a common identifier that is global in scope, data within and across databases can be accurately linked, thus providing the infrastructure for significantly improved name searching. Moreover, by also sharing the metadata resources that further describe the unique identification, ISNI participants are cooperating to achieve high quality data and at the same time realize processing efficiencies. Rights management societies have an additional constraint in that the data provided to them is in the most part bound by legal confidentiality clauses and is linked to sensitive payment information. ISNI, for them, enables linking into other databases via a neutral identifier that is one step removed and enables other public domain data, in particular VIAF information, to be used as the publicly facing metadata.

The International Confederation of Societies of Authors and Composers (CISAC), the International Federation of Reproduction Rights Organisations (IFRRO), the International Performers’ Database Association (IPDA), ProQuest, OCLC, and the Conference of European National Librarians (represented by Bibliothèque Nationale de France and the British Library) met over the period of three years, firstly as participants on the working group that developed the ISNI standard and subsequently for founding the ISNI International Agency (ISNI-IA), which will administer the ISNI assignments and registration. The ISNI-IA, an unprecedented cross-domain alliance, was incorporated in
London on December 22, 2010. The consortium members will bring together data from more than 300 rights management societies and 26,000 libraries worldwide for the initial ISNI implementation.

**ISNI Architecture**

ISNI cannot be administered in the same way as other resource-oriented identifiers such as the ISBN, where publishers are pre-allocated ranges of identifiers which they progressively apply to their new publications. Creators are not bound to any one publisher or distribution outlet or to any one domain and they publish and collaborate across national borders. Singers are often also composers and may write autobiographies or other books; as an example, Paul McCartney has written poems and a book for children. Thus, allocation and administration of ISNI needs to be controlled centrally in order to avoid, as much as possible, duplicate identifiers for any one creator.

Yet, a fully centralized system would be unmanageable on a global scale. The task of collecting the data, ensuring its quality and completeness, and then disambiguating it will be better done with a global network of participating organizations. It is also important that the assigned identifiers be diffused as widely as possible among databases and indexes accessible over the Internet to facilitate the correct linking and exchange of data. After careful deliberation, the system eventually settled on was a centralized database for requesting and referencing the ISNI identifiers with an international network of registration agencies, responsible for collecting high quality metadata and sharing in the tasks of disambiguation and diffusion. In addition to the registration agencies, some organizations will make their systems and databases available to the central system for verification and reference, as illustrated in Figure 1.

**The Initial Database**

In order for ISNI to reach its potential, industry-wide use is a key factor for encouraging adoption by research communities and internet databases generally. The ISNI-IA will ready the ISNI for industry adoption by first creating the initial ISNI database, which will allocate ISNI identifiers by processing data from the databases of the consortium’s founding members and affiliate organizations. The ISNI assignment system will be launched in the third quarter of 2011 with an initial database of assigned ISNIs. The base cross-domain file for ISNI is VIAF, the Virtual International Authority file, created over the last six years by combining the authority files of 19 major sources, mostly national libraries, and including the NACO/LC file, an international cooperative library name authority file representing over 300 libraries worldwide. VIAF currently contains 14 million names.

**Name Disambiguation**

Confidence and quality are being emphasized in the creation of the ISNI database. The matching techniques of VIAF have been adapted and employed in the ISNI system. Data files from the founding members of the ISNI-IA are progressively being matched against VIAF. ISNIs are being allocated where
there are more than three VIAF sources or two independent sources. Each allocated ISNI is assigned a confidence level in the data itself and another for the level of confidence in the matching. Thus, high confidence is placed on matching data, collected independently where one of the sources has direct contact with the identity, as is the case with a royalty management society.

The amount of data required for distinguishing one identity from another varies by the amount a name has been used and by the number of name variants. Core ISNI metadata consists of:

- Name of public identity (e.g., surname, forename, prefix, suffix)
- Name variants (e.g., Bernard Shaw, George Bernard Shaw, G.B. Shaw, and the name in a non-Roman script, but not including pseudonyms that are considered different identities)
- Creation class or classes
- Creation role or roles
- URI to contributing source or sources

Yet this data is frequently not sufficient for disambiguation. Many other data elements are collected that play a role in differentiation such as co-authors, affiliations, and publishers, but the most significant are the titles of works in which the identity has played a role in the creation, and birth and death dates. For institutions, other elements are more important, e.g., geographic location.

Where there is any doubt, multiple records for potentially the same identity are being kept in the database for manual review and an ISNI is not assigned until disambiguation is finalized. Once an ISNI is made available for widespread use and diffusion, it is hard to change or correct it. If two identifiers have been issued for the same identity, mapping from a deprecated ISNI to a correct one can be communicated and the ISNI reference database can be consulted, which will accept both the correct and deprecated ISNI and point to the correct metadata. It is considerably more difficult to diffuse corrections where one identifier needs to be split into two. The original ISNI needs to be deprecated and then point to two different identifiers, such that human review is necessary to select the correct metadata.

One of the challenges is to unite data from separate data sources where different criteria are favored for disambiguation. VIAF, for example, uses resource titles and, to a lesser extent, dates. How can you then tell one identity from another? For example Will Smith, born 25 September 1968 is an American actor, film producer, and pop rapper who performed under the pseudonym The Fresh Prince. But he is not the Will Smith (born 1971) who wrote “How to be cool” (see Figure 2).

The assignment system works best in an online, interactive mode where multiple records can be presented and either the best chosen or proof given that a separate identity is involved. Here the registration agencies (RAGs) will play an essential role in assuring the completeness and quality of the request data before it is submitted and assisting in disambiguation by providing additional information as demanded in the system’s response. Consultation of external sources will play a role. Also important is the need for the creators themselves to be able to provide input to their
There are two data formats for making requests: an XML schema and a simpler tab delimited format. The latter is envisaged for simple mapping from existing databases and submission of a file for bulk processing.

Requests will come into the system via the RAGs. At least one RAG will make a web-based individual request system, open to all, with a charge in accordance with the RAND principle, mentioned above. There are two data formats for making requests: an XML schema (see Figure 3) and a simpler tab delimited format. The latter is envisaged for simple mapping from existing databases and submission of a file for bulk processing. With bulk processing, 100% assignment cannot be guaranteed. There will always be a residue requiring further analysis. Requests using the XML request schema can be submitted either in bulk or interactively using the Atom Publishing Protocol.

Adoption and Relationship with Other Initiatives
ISNIs can be assigned to all entities that create, produce, manage, distribute, or collaborate in creative content including human beings (alive or not), legal entities (such as academic institutions, publishers, and societies), or fictional characters. The scope of ISNI is broad, though ISNI’s initial database will only include personal names. Institutions are also in scope and will be included thereafter. NISO’s
An Institutional Identifier committee has produced a set of metadata for institutions and is recommending adoption of ISNI as the identifier scheme for institutions in the supply chain. (See the F and ISNI article on page 26.)

The European Arrow Project (Accessible Registries of Rights Information and Orphan Works towards Europeana) is a consortium of national libraries, publishers, and collective management organizations that is supported by the European Commission. The consortium favors the use of ISNIs in conjunction with the International Standard Text Code (ISTC) as the fundamental building blocks for rights management administration.

Within the music industry, there is a lot of interest in ISNI, in particular by the record labels wishing to provide ISNIs to accompany their data submissions to multiple exposure and distribution sites. This industry is expected to be an early adopter of ISNI for both performers and composers.

ORCID is an initiative that is seeking to disambiguate researchers and writers of articles in scholarly journals. (See related article on page 10.) Their scope is a subset of the scope of ISNI and there has been communication between the two groups on the possibility of using the same identifier scheme. ISNI's initial database is to include two files from JISC (see Names project article on page 14): names from the Merit project, which provides names of researchers covered in the 2008 Research Assessment Exercise (RAE) data, and Zetoc, a file of authors of theses from ProQuest and ProQuest's Scholar Universe. It would be confusing for the two initiatives, ISNI and ORCID, to be producing identifiers for the same identities in the same time frame. The two systems being developed seem to be complementary, with that of ISNI focusing on name disambiguation and metadata registration and that of ORCID developing a researcher-facing system. Thom Hickey's Outgoing blog contained an entry in July 2011, VIAF and other IDs, concerning the relation of VIAF, ISNI, and ORCID and a possible pivotal role for VIAF.

**Conclusion**

The initial implementation phase of ISNI is concentrating on creating a database of assigned ISNIs of high quality and high certainty, together with a framework enabling name disambiguation that incorporates input from Registration Agencies, verification sources, and the creators themselves. doi:10.3789/isqv23n3.2011.02

---

**JANIFER GATENBY**  
@janifer.gatenby@oclc.org is EMEA Program Manager Metadata with OCLC B.V., and a member of both the ISO ISNI Working Group and the NISO Institutional Identifier (I2) Working Group.

**ANDREW MACEWAN**  
<andrew.macewan@bl.uk> is Head of Collection Processing at the British Library and a member of the ISO ISNI Working Group.
Orcid: Unique Identifiers for Authors and Contributors

A publication of the National Information Standards Organization (NISO)
Unique identifiers for authors and other contributors to scholarly works provide the basis for proper attribution. Proper attribution of contributors throughout the research ecosystem facilitates scientific discovery and impact tracking.

Researchers could use their author identifiers for manuscript submissions and grant applications, and to help locate potential collaborators. Universities could use author identifiers to help them track the research outputs of their researchers. Unique author identifiers will allow publishers to better track authors and reviewers in their journal submission systems. They also allow reliable linking to other articles by the same author, promoting the discovery of related scholarly works. Funding agencies could use unique author identifiers to facilitate the grant application process and to track the scholarly works resulting from their funding.

Despite these obvious benefits to all stakeholders, unique author identifiers have not yet seen widespread use in the research ecosystem. We currently have a fragmented system with many institutions, societies, publishers, and funding agencies using their own identifier scheme, and with author identifier systems that are either specific to a geographic region or scientific discipline. Of the commercial and nonprofit author identifier systems not limited to region or discipline, none has become a predominant player. Author identifier systems are most successful where they have been instituted on a national level, e.g. Brazil (LATTES) or the Netherlands (NARCIS).

The success of unique author identifiers depends on reaching a critical mass. This goal can only be achieved with an open and global registry for unique author identifiers. The Open Researcher & Contributor ID (ORCID) initiative was started in November 2009 to achieve this goal. ORCID has learned from other author identifier systems, and also from the experience of building the CrossRef service of digital object identifiers for scholarly works. ORCID was incorporated as a non-profit organization with a governing Board of Directors in August 2010. ORCID is open to any organization with an interest in scholarly communication and has more than 240 participating organizations as of July 2011. The first public version of the ORCID service will be launched in 2012, at which time ORCID will start charging fees to sustain the service. In this article I want to describe some of the important decisions that were made in order to ensure widespread adoption, and therefore success, of the ORCID service.

SCOPE

A number of author identifier services are focused on a particular region (e.g. LATTES for all Brazilian researchers) or discipline (e.g. RePEc for researchers in economics). Although easier to implement, this approach has important shortcomings. Many scholarly works are cross-disciplinary and/or involve researchers from different countries. This makes these focused identifiers unsuitable for many scenarios, e.g., the journal submission system of a multidisciplinary journal. The cost of implementing several author identifier services is also higher than building one system covering all disciplines and geographic regions, and many researchers will probably be reluctant to maintain more than a few author profiles.

ORCID not only transcends discipline, national, and institutional boundaries, but will also interact with existing scholarly author identification systems. The challenge to build a global system of unique author identifiers is not so much a technical one, but rather the very difficult task of coordinating a very large group of stakeholders.

Some identifier services are even broader in scope than ORCID. The International Standard Name Identifier (ISNI) aims to assign a unique identifier to all creators of digital content. OpenID is a decentralized identifier mainly used to sign in to multiple websites with the same account. The main focus of these and other services is not the proper attribution of scholarly works. OpenID is an authentication mechanism, and could be used in combination with ORCID or other author identifiers. ISNI will also serve authors, actors, publishers, and performers; the scholarly researcher is not the main focus of ISNI.

CONSENT

Unique author identifiers are fundamentally different from digital identifiers for objects. Assigning a unique identifier to a living person and using this identifier in describing this person are obviously very personal matters. It is important that authors are involved in this process, otherwise we risk that they will not accept and use the author identifier service. Before creating a unique author identifier, that author should have consented in creating this identifier. This is best done in a system where authors can apply for a unique identifier themselves. A reasonable alternative is a system where authors give permission that unique author identifiers are created on their behalf, usually by their employer. The ORCID service will allow both options.
When unique author identifiers are used in the research ecosystem, consent of the author should be obtained wherever this is possible with reasonable effort. This consent could be informal, e.g., asking authors to provide their unique author identifier when submitting a manuscript. Consent is obviously not an issue for deceased authors and could be handled differently for authors that are no longer active.

**OPENNESS**

A registry for unique author identifiers can only become a widely adopted standard if it doesn’t set any barriers for participation. Participation in ORCID is open to any organization that has an interest in scholarly communications. All researchers will be able to create, edit, and maintain an ORCID ID and profile free of charge. All profile data contributed to ORCID by researchers or claimed by them will be updated once a year and available for free download in standard formats (subject to the researchers’ own privacy settings), released under the Creative Commons Zero waiver. A survey among ORCID participants in October 2010 demonstrated the need for ORCID profile data to be freely and openly accessible.

Openness is not limited to participation and profile data. All software developed by ORCID will be publicly released under an Open Source Software license approved by the Open Source Initiative. For the software it adopts, ORCID will prefer open source. Using open source software will facilitate participation in the development of the ORCID service and will make it more attractive for third parties to develop software that integrates with the ORCID system.

**TRUST**

A unique author identifier service will only become popular if the service is trusted. Researchers would be reluctant to enter biographic and bibliographic information into a service that doesn’t give them control over the privacy settings of the information collected by them and about them. The ORCID service will allow researchers to make profile data not only public or private, but also to share private information with selected organizations, e.g., during the grant application process.

The first version of the ORCID service will rely on researchers claiming their own identifiers, but will also allow institutions to create and maintain identifiers for their researchers. Because of this self-claiming, ORCID will initially focus on currently active researchers. Users of an author identifier system have to trust the claims made in an author profile, and this is not possible in a system that relies solely on self-claims.

It is envisioned that later versions of the ORCID service will allow external claims. Multiple external claims will increase trust into the data provided by ORCID, but will also increase the complexity of the service. ORCID will not try to create one unique record from all claims about a particular author, but instead will link these records together, maintaining the provenance and avoiding issues with conflicting claims. To track the provenance of every external claim, ORCID will use unique identifiers for organizations.

Both individual researchers and organizations will have more trust in an author identifier service that is not controlled by a single organization, whether commercial, nonprofit, or government. ORCID is therefore set up as a nonprofit organization that is governed by representatives from a broad cross-section of stakeholders, the majority of whom are not-for-profit.

**REPUTATION**

A unique author identifier in itself has limited value; we have to associate the identifier with biographic and bibliographic information. This information could be in the form of digital object identifiers (DOIs) for journal articles or research datasets, other digital identifiers for scholarly content, or could be free-form text, e.g., for research grants and awards. With this information we are building an author profile, and this profile can be maintained in the author identifier service and/or in external services.
All author identifier systems collect profile information, but many of them are limited to publications. The first version of the ORCID service will collect past and present institutional affiliations and citations to scholarly papers. Future versions of the ORCID service will also collect other scholarly works, and this could include conference abstracts, research datasets, patents, and other scholarly activities. From the start, the ORCID identifiers will be available to other services via an API. Information about journal articles will be directly imported via the CrossRef service, and it is expected that tight integration of the ORCID service with journal submission systems, institutional repositories, grant submission systems, etc. will associate authors on an ongoing basis with their scholarly works at minimum burden to researchers.

ORCID author profiles can be used for knowledge discovery and for academic metrics. Making a substantial number of author profiles available to the scholarly community will create valuable new opportunities and is an important goal of the ORCID organization. ORCID is looking into providing some of the aggregated profile information (which is made available according to author privacy settings) as a potential revenue source.

**PERSISTENCE**

A unique author identifier service will only become widely used if the organization behind it is sustainable and can guarantee the longevity of the service. The ORCID organization will be run by a small number of staff and is relying on voluntary work from participating organizations and the use of open source software wherever possible. Despite this, building and maintaining a global author identifier service is still expensive, and it is important to understand that ORCID will have to charge fees to ensure the sustainability of ORCID as a not-for-profit, charitable organization. The ORCID business model will be finalized by the end of 2011, and will most likely include a combination of membership fees, fee-for-services, and sponsorship. (The service will always be free to use for individual authors.)

ORCID is applying for 501(c)(3) tax-exempt status and this requires it to make provisions for passing on the responsibilities to a similar organization should ORCID cease to exist. A focus on persistence requires ORCID to control the intellectual property and the software running the service.

The unique author identifier itself also has to be persistent and ORCID identifiers can’t be removed from the system once they have been issued. For privacy reasons, ORCID allows researchers to decide which of their data they want to make publicly available. All the public data will be made available under a Creative Commons Zero waiver, updated at least once a year. This policy encourages the regular distribution of all public data, decreasing the likelihood that data are lost. This process also holds ORCID accountable, as the system could be duplicated somewhere else if ORCID would take a direction not supported by the community.

ORCID is taking all the necessary steps to ensure widespread adoption of the service. Once the service launches and unique identifiers for authors and contributors become as common as digital identifiers for scholarly objects, it will be hard to think back to a time when it was impossible to do all the things made possible by unique author identifiers.

**RELEVANT LINKS**

- Creative Commons Zero (CC0): creativecommons.org/about/cc0
- CrossRef: www.crossref.org
- ISNI: www.isni.org
- LATTES: lattes.cnpq.br
- NARCIS: www.narcis.info
- OpenID: openid.net
- Open Source Initiative: www.opensource.org/
- ORCID website: www.orcid.org
- ORCID Principles: www.orcid.org/principles
- ORCID FAQ: www.orcid.org/faq
- RePEc: repec.org
The Names Project: A New Approach to Name Authority

ALAN DANSKIN, AMANDA HILL, AND DANIEL NEEDHAM

Unique, unambiguous identification of researchers has become a hot topic in recent years, with a number of initiatives now under way to solve the problem. It is a well-known fact that personal names are not sufficient as a means of distinguishing between individuals: there may be more than one person with the same name and, if only initials are used (as is the case in many bibliographic databases), this problem is compounded.

A person’s name may also change during the course of his or her lifetime, or be represented in subtly different ways, creating a need to link all those alternative forms of name together in order to be sure that the materials produced by that individual can be reliably identified and (if desired) collected together.

The Names Project was funded in 2007 as part of the JISC’s Repositories and Preservation Programme. It had been recognized that:

Searching by authors’ names has been among the top search methods by repository users. When a repository grows to substantial size, it is often the case that name variants cause headaches for both the users and repository managers.[Xca]

When the contents of different repositories are aggregated, these problems in retrieving all (and only) relevant materials are compounded. The JISC’s call for proposals in September 2006 therefore had a specific requirement for a project which would investigate:

…the potential for the development of a Name Authority Service and factual authority for digital repositories, to support cataloguing, metadata creation and resource discovery in the repository environment.[JISC]

A joint bid submitted by Mimas (at The University of Manchester) and the British Library was successful and work began on the project in July 2007. Early activities included identifying the requirements of the repository community and reviewing the work of existing projects and services in the field of researcher identification and name authority. National libraries have been creating name authority files for authors of books for many years, starting with card catalogs and now maintaining electronic files in MARC format. However, authority files for the creators of journal articles
and other electronic resources often do not exist in library systems. The increasing use of subject-based and institutional repositories to hold working papers, reports, research data, and pre-refereed and post-refereed versions of articles has led to a corresponding rise in the number of authors identified in such systems.

Assessment of research activity has been a significant part of life for researchers employed in the UK’s universities since the introduction of the first “research selectivity exercise” in 1985-86.[Day] Research Assessment Exercises (RAEs) have been carried out in 1989, 1992, 1996, 2001, and 2008, with peer-reviewed analysis of publications forming an important part of the process. It seems likely that measurements of the impact of research will be taken into account in the RAE’s successor, the Research Excellence Framework (REF). With the increased availability of research materials in repositories and other online locations, it is becoming vital that researchers are reliably associated with their publications in order to show how the outputs of a particular individual are being used and affecting the work of others. Uniquely identifying researchers would assist in this process.

**View from Mimas**

Mimas is a national data center based at The University of Manchester. The department has a history of providing innovative ways of connecting users with research information and developing technological infrastructure services to support UK academic researchers. Finding a solution to the researcher-identification problem would help to support these objectives.

One of the initial aims and ongoing activities of the project has been the design and development of a prototype name authority service for individuals and institutions in order to demonstrate the feasibility of such a system. The original design of the prototype was based upon the project’s landscape report, stakeholder requirements-gathering exercises, and consultation with the developer community. Subsequent development of the prototype has been an iterative process, due to the dynamic environment within which it needs to fit and the changing requirements of the varying stakeholders.

The prototype was envisaged as a piece of middleware, comprising of a store of name authority records created using a data model designed by the British Library, and an API through which the records could be queried by external services. In order for the service to be viable, the first thing that was required was a large data set around which the service could be built.

With no available pre-existing set of data pertaining to individuals and their identities, it was necessary to build our own records from scratch. Two approaches to building such a data set were identified: firstly by acquiring access to external data sources containing information relating to individuals and institutions and attempting to automatically disambiguate the unique entities within them, and secondly by providing functionality for individuals and institutions to contribute data to the service themselves. It was determined that the former approach would be more appropriate in the initial phase of prototype development, with the latter being introduced at a later stage.

The prototype was designed to be as flexible as possible in how it acquires and processes external data, which would be accessible in varying formats and provide diverse types of information. For this reason the disambiguation side of the prototype comprises two logical sections. The first is a collection of data source handlers, each of which is tailored to the specific external data source it relates to.

**TWO APPROACHES TO BUILDING A DATA SET:**

1. Acquiring access to external data sources containing information relating to individuals and institutions and attempting to automatically disambiguate the unique entities within them

2. Providing functionality for individuals and institutions to contribute data to the service themselves

CONTINUED »
This provides the functionality to pull in data from the specified source and then convert it into a collection of internal Names record objects. Data source handlers can be added each time a new external source is acquired. The second logical component accepts these converted Names record objects and attempts to identify and disambiguate the unique entities contained within the source data, first against other records from that source (if necessary), and then against any existing Names records containing information derived from a number of different sources that have the potential to match.

In order to analyze entities identified in different records for potential matches, a disambiguation algorithm examines the various attributes of the entities in question. By using pre-configurable thresholds for the differing matching criteria, the algorithm attempts to determine whether or not the compared entities match according to these rules. Where no match is found, a new record is created for that entity within our own database and assigned a unique persistent identifier. Where a match is found with an existing Names record, that record is updated with any new information. Some of the attributes which have proved most useful for matching, aside from entity names, include collaborative relationships, publication title keywords, fields of interest, and institutional affiliation.

Where a match is found with an existing Names record, that record is updated with any new information. Some of the attributes which have proved most useful for matching, aside from entity names, include collaborative relationships, publication title keywords, fields of interest, and institutional affiliation.

Using this data. Whilst initial results from the disambiguation process seemed promising, scaling up management and testing of such a large data set proved problematic given the resources we were working with.

Consequently, we began to look at alternative data sources including MERIT data from the 2008 RAE. The JISC-funded MERIT project worked on cleaning up the information submitted to the RAE and the resulting data set, containing information on 45,000 of the UK’s top researchers, was a more manageable size to work with. After creating a data source handler to process the MERIT data, we successfully disambiguated the entire data set with a very high level of accuracy, eventually eliminating all mismatched records.

In May 2011, following a period of quality assurance carried out by the British Library (described in more detail below), the Names records derived from MERIT were made permanent, and an export of our data was sent to the International Standard Name Identifier (ISNI) initiative, to be matched against their records.

Subsequent to establishing a base set of Names records from the MERIT data, we are now looking at matching the derived records against a subset of the Zetoc data. We are also attempting to match against data exposed by the RDF output of the EPrints repository software, using The University of Southampton’s repository as a test case.

In order to provide access to the Names records that had been created during the disambiguation process, it was necessary to design and implement an API through which queries could be made. The API needed to provide facilities to search over the Names records and return results in a flexible way to meet the varying requirements of the stakeholders. Initially SOAP was chosen as the means of providing data search and retrieval; however, following feedback from the JISC developer community during the initial stages of the project, it was decided that a RESTful approach would be more appropriate for the audience that would be using it.
Each Names record is assigned a unique identifier, and this identifier is a resolvable URL as part of the API. The API also provides functionality to search for records using a variety of criteria and returns the results with differing levels of detail and in different output formats. Consequently the API provides a robust and flexible method for searching Names records and can easily be integrated into external systems with minimal effort. This has been demonstrated in both a search interface developed for testing and demonstration of the capabilities of the prototype, as well as an example tool that was written to illustrate the way that external services could use the API to autocomplete a name field in a form with Names data.

The development of the prototype will be ongoing, with the aim of increasing the quality and quantity of records as well as the functionality that the API provides for interacting with them. As part of this work we will be looking at acquiring new data sources to process, as well as refining our disambiguation algorithm to increase the accuracy of results. We will also be reviewing the API and working with external services and repositories to facilitate integration between their applications and the prototype.

View from the British Library
The British Library is the national library of the United Kingdom and is one of the six UK and Irish libraries entitled to receive UK publications under legal deposit. The British Library is a partner in the Names project because control of the names of authors and other contributors to publications is an important and expensive element of cataloging items for the collection.

One of the functions of the library catalog is to enable a user to find “all resources associated with a given person, family or corporate body.” [IFLA] To satisfy this requirement it is necessary to identify each individual entity uniquely and to provide links between the variant forms of names by which they are known. In a library context, these functions of identification, disambiguation, and linking are provided by the name authority file. In current cataloging practice the focus is on disambiguation of entries (headings) in a browse index. In a web context, where library metadata has to mix with metadata from other domains, entities have to be explicitly identified to enable joined up services. The Names project has engaged with initiatives developing international identifiers for researchers, including the ORCID Initiative and ISNI.

The way in which authority control is done by libraries is challenged by audience expectations and by the volume of resources that will require authority control. The focus on controlling the authors of printed books no longer satisfies the needs of researchers, who want journal articles, conference papers, data sets, pre-prints, and other resources. The number of new books published in the UK and received by the British Library through legal deposit is about 130,000 per annum; the number of journal articles added to the ETOC system is approximately 2.5 million per annum. Manual processes are not scalable to meet this demand. Automation or semi-automation of authority control processes would enable the British Library to identify individuals in ETOC records and link these identities to existing authority files.

The library has contributed its expertise and metadata to the Names project by contributing to development of the data model, specifying mappings to output formats, and testing of samples of metadata disambiguated by Names.

Testing has been conducted in three main phases, described in more detail below. All of the testing done at the British Library involved evaluation of sample data by catalogers following normal practices used by authority control staff to identify and disambiguate individuals of the same name. Catalogers consulted external sources, predominantly institutional or personal websites, to confirm that identifications made by Names are secure and accurate. The manual review by British Library supplemented extensive testing and validation carried out by Mimas.
Data analysis

The ETOC data numbers approximately 38 million article records. To gain an understanding of the data, samples were extracted for specific personal names. Selecting specific names meant that levels of duplication and disambiguation could be evaluated. The overlap with the LC/NACO authority file was also evaluated and the low incidence of matching influenced the decision to defer loading LC/NACO to Names.

The findings confirmed assumptions about the ambiguity of the data. For example, the name “Birtwhistle, G.” concealed seven different identities. The analysis also confirmed expectations that there would be many entries for the same person. For example, “Birtwhistle, G. #2” was associated with sixteen different article records. The low frequency of matching with NACO confirmed that articles receive very little attention from authority control catalogers in other institutions and disambiguation would have value beyond the British Library.

Evaluation of disambiguation

A sample file was prepared to enable comparison of automated outputs with manual authority control. The sample consisted of 375 article records associated with the name C. Abbot. Manual review of the sample took approximately four weeks.

The manual review process was time consuming, but very valuable. It highlighted the limitations of the ETOC data for matching, but more positively a problem with the weighting given to subject classification numbers (Dewey Decimal Classification) was identified, which when adjusted improved the results.
Disambiguation of MERIT data

<table>
<thead>
<tr>
<th>Source file size</th>
<th>45,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Potential Mismatches</td>
<td>33</td>
</tr>
<tr>
<td>Actual Mismatches</td>
<td>2</td>
</tr>
<tr>
<td>Non-Matches</td>
<td>2257</td>
</tr>
<tr>
<td>Non-Matches reviewed</td>
<td>227</td>
</tr>
<tr>
<td>Actual Matches identified</td>
<td>7</td>
</tr>
</tbody>
</table>

The MERIT data is a much smaller and better controlled set than the article records and includes institutional affiliations and researcher IDs, neither of which is present in the ETOC records. For this phase of review, the results were filtered to identify possible mismatches or non-matches. Mismatches occur when names of two different individuals are matched incorrectly. The “non-matches” were records for which several individuals with the same or very similar name were identified, but not matched.

There were too many non-matches to review them all manually; therefore every 10th result was reviewed. Mismatches are considered to be more serious than non-matches and all the potential mismatches were reviewed. Only two genuine mismatches were identified. In one of these cases the individuals concerned turned out to be twins who worked at different institutions but had the same initial and family name and were associated with the same paper. One result of this discovery has been the adjustment of the algorithms to prevent matches between individuals with the same name who have collaborated on the same paper.

The MERIT data, deduplicated and disambiguated by Names, provides a core of reliable identities for UK researchers and academics against which other data sets, such as article records, can be matched. The MERIT records have been exported as the first contribution by names to the ISNI database.

Manual review of the Names outputs highlighted the importance of human inspection of the results. Sampling and filtering created manageable workloads for catalogers, cutting the time to review representative samples from weeks to days. Input from the reviewers has improved the matching and disambiguation algorithms. Future services will require an element of human review to resolve ambiguities and for quality assurance.

Conclusion

Unique identification of researchers is an area of intense interest in the UK and beyond. The Names Project team has aimed to test the feasibility of a service that would provide disambiguation and identification of researchers and make the resulting records available to the wider research community. The work of the project has produced a core set of disambiguated researcher identifiers, accessible through a flexible API, which could be used as the basis of a future name authority service. Plans for further enhancement of the service would include allowing the researchers themselves (or their representatives) to supply information that would improve the accuracy of the data in the system, and further collaboration with related international initiatives such as ORCID and ISNI.

ALAN DANSKIN <alan.danskin@bl.uk> is Metadata Standards Manager at The British Library. AMANDA HILL <amanda@hillbraith.com> is an Archival Consultant who manages the Names project on behalf of Mimas at the University of Manchester. DANIEL NEEDHAM <daniel.needham@manchester.ac.uk> is Technical Officer for Mimas.
Identify This! Identifiers and Trust

I am a self-confessed identifier-dweeb. And I am acutely aware that talking about identifiers is just about the best way to shut down conversation at all but the geekiest of social gatherings. Identifiers are boring. Talk about identifiers and people think about inventory control, supply chain management, rollup reports, and government bureaucracy.

This is because most people are only familiar with the more prosaic applications of identifiers. Yes, ORCID will make manuscript tracking and researcher evaluation systems more efficient. Yes, Institutional Identifiers will allow publishers and librarians to manage subscriptions more reliably, and yes, CrossRef DOIs enable publishers to avoid having to negotiate complex bilateral linking agreements—but these quotidian applications of identifiers often obscure their more profound and long-term importance. Identifiers are the foundation upon which we will increasingly rely in order to publish trustworthy electronic content. In short, identifiers are about “trust.”

Certainly anybody having to go through the daily slog of weeding out spam, avoiding phishing attacks, or running anti-virus software is aware of the trust problem on the internet. And this is to say nothing of the relatively new problems associated with detecting astroturfing, link-farming, sock-puppets, or any of the myriad of increasingly sophisticated techniques that unscrupulous people are using to promote their content and agendas.

When Tim Berners Lee and Nigel Shadbolt recently launched an initiative to create an academic discipline called “Web Science,” they summarized the issue of trust and the web as follows: “How can we determine whether we can trust the material emanating from a site? The Web was originally conceived as a tool for researchers who trusted one another implicitly; strong models of security were not built in. We have been living with the consequences ever since. As a result, substantial research should be devoted to engineering layers of trust and provenance into Web interactions.”

Indeed, but what do we mean by “trust?” Phil Windley in his book, Digital Identity, defines trust as: “...a firm belief in the veracity, good faith, and honesty of another party, with respect to a transaction that involves some risk.”

One of the techniques that we normally use for evaluating trustworthiness is to assess the provenance of the entity we are being asked to trust. “Do I know this person?” “Am I familiar with this institution?” In “meatspace” we have countless cues and heuristic tools that we automatically use for evaluating trustworthiness. “Do I recognize this person’s face and/or voice?” “Have I seen other branches of this store before?”

Similarly, with physical media, we could use heuristics as an aid in judging the trustworthiness of the content therein. The binding of the content, the weight of the paper and the quality of the printing, the presence/absence of scholarly apparatus (footnotes, indexes, bibliographies, graphs, equations, etc.)—all gave us clues as to the reliability and authority of the content in question.
But how do we judge trustworthiness on the Internet—a world where content is protean, provenance is vague, and identity is cheap? “Do I believe this e-mail is who it says it’s from?” “Is this web login page really from my bank?” “Are the blog reviews of this restaurant authentic or were they written by shills?” We don’t really have good tools for answering these questions. The cues and the heuristics that we can use for such evaluations on the Internet are negligible compared to the counterparts that we use every day in our physical interactions.

In my particular industry, scholarly publication and communication, the issue of trustworthiness is paramount and Windley’s definition of trust is useful because it touches on many salient aspects of the scholarly publication process. In the case of a researcher, the “transaction” we are talking about is that of consuming and acting on formally or informally published information. The “risk” associated with this transaction is, minimally, that the researcher wastes time reading or acting on information that is somehow flawed. But often the risk can be far higher; it can, for instance, damage one’s reputation or do serious harm to third parties.

So how do researchers mitigate this risk? Ultimately, of course, researchers use their discipline expertise to assess the content before they act on it. However, before assessing the content (a time-consuming process), the researcher often employs a useful heuristic shortcut, that is they look at whatever “brands” are associated with said content. Recognition of the brand will often tell the researcher something about the risk they are taking in consuming and using the information. Of course, the degree to which the brand can serve as a shortcut varies greatly from brand to brand and it is this level where publishers fiercely compete to earn the researcher’s recognition and, one hopes, trust.

Similarly, the researcher understands that the use of brand as a shortcut is a heuristic and like all heuristics, it is fallible. The best journals occasionally publish rubbish. Unknown journals occasionally publish gems. There are also times when researchers cannot use brand as a shortcut. They may be unfamiliar with the brand because they are not experts in the field (e.g., a novice researcher, a cross-disciplinary researcher, a journalist, or a government functionary), but it might be because the brand has not yet been established. Even the most powerful publishing brands were, at some point in their history, entirely unrecognized. When researchers cannot use brand as a shortcut, their next step at attempting to identify reliable information is to establish the provenance of said content. To do this, the researcher gathers and confirms evidence as to the time and place the content was created, evidence relating to the parties responsible for the creation and

CONTINUED »
production of said content, and evidence of the procedures that were used to ensure the content’s integrity. Researchers expect the scholarly record to aid them in assessing the provenance of content. And let us remember, this scholarly record may go back years, decades, or even centuries.

It is here where identifiers have become increasingly important in establishing the provenance and trustworthiness of electronic content. People’s names “collide.” People change their names and sometimes just record their names differently according to mood or situation. Organizations—not just companies, but universities, government departments, and entire countries—mutate, merge, split, and sometimes disappear. Content is increasingly dynamic, increasingly copyable, and increasingly modifiable. All of which, combined with the general internet trust issues discussed above, means that the apparently simple act of accurately citing and crediting scholarly work is becoming more fraught. A researcher has a reasonable expectation that when he or she cites something today, that another researcher in twenty years time who follows that citation will see exactly what was cited, not some new or modified version of what was cited. And similarly, researchers expect that their work will be credited to them properly and not to somebody who shares their name.

For this to work accurately and to scale, we will increasingly have to rely on unique identifiers for people, organizations, and content. So you see, even though the value of identifiers in the short term might be to make our operations more efficient, in the end identifiers will become the foundation of a new epistemic infrastructure for reliable and trustworthy computer-mediated communication. Identifiers are not boring. Identifiers are about trust.

GEORGE BALDWIN <gbilder@crossref.org> is Director of Strategic Initiatives at Crossref.
In current discussions of metadata in the publishing supply chain, much attention is necessarily paid to issues surrounding title metadata and the effective use of the ISBN. Other standards, such as the ISTC (International Standard Text Code) and DOI® (Digital Object Identifier), are mentioned in this regard as well. Often overlooked is the role played in the industry supply chain by that less glamorous cousin to the ISBN, the SAN (Standard Address Number).

SAN was initially developed at R.R. Bowker in response to retailers’ requests for a standardized account to facilitate ordering across the supply chain. As the ISBN is the standard ordering number for the book publishing supply chain, the SAN was intended to meet the need for an accurate way to contact trading partners so that orders could be processed effectively, accurately, and seamlessly. In the end, rather than becoming an account numbering system, though, the SAN system was set up to identify addresses, the key piece of information for accurate business transactions.

The SAN is a unique identification code for each address of an organization in the publishing supply chain. The SAN enables product discoverability by facilitating standard daily business activities, such as order placement. A simple example of the use of the SAN is as a means for publishers to identify retailers and link their ordering processes with them. Another example is that of a bookstore chain using SAN to identify publisher shipping addresses for returns.

In the United States alone, more than a hundred thousand organizations are involved in repetitive transactions with one another in the publishing industry supply chain. These transactions include purchasing, billing, paying, crediting, refunding, lending, exchanging, shipping, and receiving. The organizations engaged in these
transactions include not only publishers and retailers, but also wholesalers, distributors, printers, paper vendors, binderies, and other vendors and suppliers to these same organizations. In addition, although not directly in the publishing industry, libraries, library binders, serial vendors, school systems, colleges, and technical institutes also take part in such transactions, since they are involved in publication purchasing and shipping.

Many of these organizations have different addresses to which bills should be sent and which differ from shipping addresses. Some organizations have multiple addresses. A printing company may very well have an ordering address, a billing address, and several warehouse addresses. Identifying which address is appropriate for a specific transaction can be confusing and challenging.

Identifying the correct point of contact among similarly named organizations can also further complicate the successful processing of transactions. For example, there are more than fifty U.S. publishers trading under the name Lighthouse Publications or Lighthouse Publishing. Failing to identify the correct, similarly-named organization provides opportunity for error, delay in billing and payment, as well as the possibility of shipments gone astray. Maintaining extensive records both for billing and tracking shipments can be complex under such circumstances. EDI systems (Electronic Data Interchange) add a further dimension to the picture.

Like the ISBN, the SAN makes sense of the decentralized nature of the publishing industry supply chain. Organizations in the publishing industry can communicate with one another more effectively by using not only the ISBN but the SAN as well. While the ISBN enables product identification, the SAN enables communication from organization to organization, connecting buyer and seller directly, whether through traditional ordering and shipping mechanisms or through electronic systems. The SAN standard accomplishes this by providing a one-to-one correspondence between SAN and the address to which it is assigned.

Approved in 1979 as ANSI/NISO Z39.43-1980 and later updated in 1993, the SAN consists of six digits, plus a modulus-eleven check digit. This seven-digit code is hyphenated after the third digit for easy human transcription, but the hyphen need not be retained in electronic systems. Each SAN identifies a single address. An organization with multiple addresses would use multiple SANs, each identifying a different one of the organization’s addresses. When an organization relocates, the SAN is updated with the new address. If an organization closes a location, the SAN cannot be reused. If an organization, such as an independent bookstore, changes ownership, the SAN can be transferred to the new organization. A new SAN need not be obtained because the address remains the same and the SAN primarily identifies the address of the organization rather than the organization itself, although the organization name does appear as part of the SAN record.

The role of the SAN as an address identifier has distinct and clearly defined functionalities. The SAN is never associated with a specific product because it is the ISBN that acts as the ordering or tracking number for a specific product. Nor can the SAN be encoded on a barcode. Despite its role in supporting accurate shipping, the SAN is distinct from shipping labels and product labels and does not appear on such carton marking.

In discussions on standards, the SAN would seem to be similar to, and is frequently compared to, the GLN (Global Location Number) assigned by GS1. However, the SAN differs as an identifier from the GLN. The GLN is broader in its application, identifying addresses and functions in order processing and shipping similar to the SAN, but is more widely applied as a location identifier, including the recognition of individual legal entities. The GLN can be modified to identify specific
locations within a warehouse, for example, and can be as granular in its location identification as a single store shelf. The GLN is not industry specific and can be used to identify any location of an object or product, such as a hospital bed or a shipment of sneakers.

The SAN, on the other hand, is industry specific, used only in the publishing industry supply chain, and is more clearly defined in its functions, thereby making it more functional and efficient for its intended use. The SAN is strictly a standard identification number for an address associated with the publishing industry, whether used for placing orders, shipping from place to place, or other common business transactions. It can be placed on any correspondence between trading partners necessitating address accuracy, such as invoices and receipts. In back-office databases the SAN can function as a link between ordering and shipping databases and as a search key for addresses in such databases. In an EDI system, such as Pubnet, the SAN functions as an electronic address in itself. Problems that can occur, such as billing errors, products shipped to the wrong points, and errors in payments and returns will almost be eliminated by using the SAN.

Significantly, when using a SAN, the need to look up account numbers as a normal and required step of order fulfillment and shipping processes ends. Without a SAN, the correct account number for processing a transaction with a particular organization must be identified before an order can be processed. With a SAN, vendors can proceed to process orders immediately without needing to locate or to assign an account number, for example. The SAN itself is used as the central number for doing business with all of an organization’s trading partners in the publishing industry supply chain. To locate trading partners’ SANs, a comprehensive database of publishing supply chain trading partners and libraries can be accessed by subscription. Bowker’s SAN-based online directory lists SAN assignments for publishers, libraries, distributors, wholesalers, bookstores, printers, book manufacturers, etc., including address and contact information.

The SAN system has escaped the intense, lively discussion and confusion over use that has recently inveigled the ISBN. This is the case for several reasons. Unlike the ISBN standard, which is evolving to meet the product identification needs of a new category of electronic publications, the essential nature of business transactions have remained unchanged so the SAN need not evolve dramatically as an identifier. The seven-digit SAN is easily associated with electronic office software, if desired, and with POS (point-of-sale) systems.

In addition, while the advent of e-books has been a boon for the self-publisher, the typical user of a SAN is a business accomplishing standard business tasks on a daily basis. The SAN meets this need because it was specifically developed for use with repetitive transactions with multiple trading partners.

Self-publishers, on the other hand, are not running a business on this scale and are often uninterested in embarking on such an enterprise. Because their business activity is minimal, they face no pressing concern to track a product or process orders among multiple trading partners. They may, in fact, be working with only one publishing services company or a single retailer. Such individuals who are publishing e-books have no need of a SAN under these circumstances.

No doubt the publishing industry supply chain will continue to transform itself in unpredictable ways. The use of identifiers will continue and, by necessity, the identifiers themselves will evolve in response to practical business needs. Whenever an organization needs to engage in ordinary transactions with multiple trading partners, the SAN will continue to be a simple, elegant way to identify the right point of contact. 

LOUISE TIMKO is Supervisor Provider Relations for Identifier Services at R.R. Bowker.

Global Location Number (GLN)  
www.gs1.org/barcodes/technical/idkeys/gln

International ISBN Agency  
www.isbn-international.org/

Pubnet  
www.pubnet.org/

R.R. Bowker  
www.bowker.com

SAN look-up database  
www.BookIndustryLocator.com

SAN Standard (ANSI/NISO Z39.43)  
www.niso.org/standards/z39-43-1993r2006/
I² and ISNI: Improving the Information Supply Chain with Standard Institutional Identifiers

The NISO Institutional Identifier (I²) Working Group was established to create a standard for an identifier that would support more efficient workflows for the information supply chain by enabling all parties—such as libraries, subscription agents, and publishers—to use the same identifier for the same institution.

For example, when a publisher receives orders for online journals from multiple subsidiary departments, it is difficult (and expensive) to make sure all orders are assigned to the correct online account of the institution. Failing to do so can result in loss of access to content, frustration with information providers, and added expense for the providers to diagnose and correct the problem. The group built on the work of the Journal Supply Chain Efficiency Improvement Pilot, which demonstrated the improved efficiencies of using an institutional identifier in the journal supply chain.

Over the past 2 years NISO’s I² working group has been seeking the best infrastructure to support a standard institutional identifier in a scalable, extensible manner. Several other standards and identifiers were studied to select those aspects which worked well and to identify unmet needs and service gaps. During this investigation, we determined that the International Standard Name Identifier (ISNI) could be leveraged to meet the infrastructure needs of I², while the needs analysis and metadata development performed by the NISO I² working group could expand the ISNI’s ability to serve institutions. This synchronicity laid the groundwork for a fruitful collaboration.

ISNI offers a solution

ISNI is an approved international standard (ISO 27729) that was developed to provide a standard identifier for people (such as authors, characters, and public figures) and the institutions with which they are affiliated (such as publishers and universities). ISNI has already received strong support from organizations in the publishing and information access arenas, and the not-for-profit ISNI International Agency (ISNI-IA) is developing the infrastructure to implement and manage the identifiers. The base of this infrastructure, the ISNI Registration Authority (RA) appointed by ISO, will be in charge of creating and maintaining the ISNI reference database, as well as the overall administration and governance of the ISNI standard. The RA’s real-time systems will allow business partners—ISNI Registration Agencies (RAGs)—to obtain a new identifier or look up a registered entry. Registration Agencies are appointed by the ISNI International Agency; any business entity with a proven interest in the scope of ISNI will be eligible for consideration. The ISNI business model encourages the diffusion and use of ISNIs once they have been established; therefore, it is reasonable to expect these identifiers to be used by organizations that are not formal registration agencies. The basic requirement is being able to handle an identifier of 16 decimal digits (the last is a check digit).
ISNI has the potential to be the solution for identification of institutions within the information supply chain. An example of how this would work is as follows:

A subscription agent becomes a registration agency (RAG).

The central registry checks to see if the institution is already registered and returns the existing identifier if it exists. If no identifier exists, the ISNI RA adds the institution to the registry and returns the newly assigned identifier.

For each new customer, the subscription agent’s system sends an automated request to the ISNI Registration Authority (RA) central registry asking for an identifier for the new customer.

The subscription agency is assured of a globally unique identifier for each customer, while the customers no longer have to manage separate identifiers for each subscription agency or other vendors with which they do business. The identifier would support collaboration, such as when multiple customer organizations form a consortium for purchase but still require individual identification. The identifiers would also support collaborative packaging and joint marketing of content by subscription agencies, while retaining individual agency identities.

The business model allows many organizations to be registration agencies and is designed to allow the identifiers to be obtained at point-of-need, while at the same time ensuring that multiple identifiers are not assigned to the same registered institution. Although there is a nominal cost for an identifier to be assigned, there is no restriction on how the identifier can be used or shared.

The core technology behind the use of ISNI as the Institutional Identifier is the simple but powerful premise of “identify once, use many” that would transform the ability of participants in the information supply chain to work efficiently, achieve economies of scale, and to innovate in an emerging digital environment.

Laying the Groundwork

Leaders of the NISO I^2 Working Group met with the ISNI Working Group to explore common interests and investigate a potential collaboration. NISO I^2’s interest is to ensure that ISNI becomes a viable standard that will be used for institutional identification and that the ISNI International Agency (ISNI-IA) infrastructure and business model extend to the identification of institutions within the information supply chain. The benefit to ISNI is the extensive exploration of institutional identification needs performed by the I^2 Working Group and the development of a strong metadata profile for institutional identification to supplement ISNI’s already solid identification of individuals.

The ISNI-IA was definitely interested but needed NISO I^2 to identify an organization (or organizations) that would become an ISNI Registration Agency (RAG) specifically in the business of registering institutions in the information supply chain. The ISNI-IA would then work with the RAG or RAGs to develop and refine their service and processes. NISO I^2 would also need to map the I^2 institutional metadata, previously defined by the Working Group, to the ISNI request schema and help develop a single schema that ISNI could use to identify both people and organizations. NISO I^2 could also play an ongoing role in facilitating the adoption of ISNI as the I^2 by helping to find other organizations to work as RAGs and by assisting with education and promotion of the identifier and its use in the information supply chain.

CONTINUED »
The I² Working Group developed an initial list of organizations that could be appropriate Registration Agencies for institutions in the information supply chain. ISNI Registration Agencies (RAGs) will have unlimited read-only access to the full ISNI database and they will be the only entities authorized to issue ISNIs.

Metadata Harmonization
An important need to facilitate collaboration was harmonization of the metadata developed by NISO I² for institutional identification with the metadata profile in the ISNI standard. Committee members analyzed the ISNI metadata in comparison with the I² metadata for equivalences, gaps, and conflicts. A comparison of the two metadata schemas revealed multiple synchronicities, which can be leveraged to map the schemas together. Janifer Gatenby as a member of both the ISNI and NISO I² working groups was able to serve as liaison to facilitate this effort and enable needed changes to be made quickly.

Two core areas of need identified by NISO I² were:

1. date ranges to specify time periods when institutional names were valid, and
2. documentation of relationships between institutions.

Date ranges are critical to support historical identification of organizations that changed their names but not their essential identities. For example, OCLC was once identified by the name Ohio College Library Center, then Online Computer Library Center, but is now simply known by the acronym OCLC. Despite the name changes, the actual organization did not change. In such a scenario, the organization would retain the same institutional identifier despite these changes, but the metadata would need to be updated. Such name changes needed to be distinguished from changes to the organizational identity, such as when two or more organizations merge into a new organization, which is documented in the “related institution” data element and would result in a new institutional identifier for the merged organization.

Documentation of relationships between institutions may be critical not only for identification purposes, but also to support business processes. Subsidiary organizations within an institution may need their own identifier for some workflows but not for others. For example, the Special Collections Department of a University Library may need its own identifier to facilitate interlibrary loans, since its policies, loan periods, and physical address may be very different from the parent library. However, for other activities, such as the purchase of materials, the parent library may be the identifier used. These relationships need to be unambiguously identified as related institutions, rather than as variant names for the same organization, to facilitate reliable and authoritative transactions. NISO I² spent considerable time developing relationship metadata and testing it in hypothetical scenarios of use to solve common problems.

Other NISO I² recommendations that ISNI is considering include the incorporation of the ISO 3166-1 standard for country codes as opposed to the use of MARC codes. One remaining issue under discussion is the ISNI metadata dependence on identification of a resource associated with the entity being identified. While resource titles are useful for disambiguating authors or publishers with similar names, institutions are not generally associated with authored resources. Therefore this requirement is generally not as meaningful or useful for institutional identification as other data elements. The I² group has formally requested modification of this requirement for institutions and believes a restructuring of the hierarchy of ISNI metadata could support this request. ISNI has already reorganized the hierarchy of some of their metadata elements to better support the identification of institutions and is considering this additional request.
Identifying ISNI Registration Agencies (RAG) for Institutions

The I² Working Group developed an initial list of organizations that could be appropriate Registration Agencies for institutions in the information supply chain. ISNI Registration Agencies (RAGs) will have unlimited read-only access to the full ISNI database and they will be the only entities authorized to issue ISNIs. I² co-chairs Grace Agnew and Oliver Pesch developed an overview of the proposed I²/ISNI relationship to assist in describing the benefits and importance of this project to the prospective institutional ISNI RAGs. The I² Working Group members then approached the organization contacts to educate them about the project, ask for support, and pave the way for future adoption.

Thus far, two organizations have expressed interest in being ISNI Registration Agencies for institutions, pending appointment by the ISNI IA. This is an excellent start, but NISO I² hopes to see additional organizations involved as RAGs to represent multiple sectors of the information supply chain, such as library and archival management systems, collaborative repositories, cloud information services, and others.

Conclusion

The NISO I² Working Group has identified a clear need for a standardized institutional identifier, clarified the necessary parameters for identification, and investigated various methods of support for implementation. The ISNI ISO standard, initially intended for individuals and characters, could provide the necessary infrastructure for support and delivery of an institutional identifier. By merging the I² and ISNI metadata schemes, the ISNI implementation could expand to incorporate the needs of robust, standardized institutional identification. NISO I² has identified two organizations that are willing to become ISNI Registration Agencies actively engaged in the identification of institutions. It is hoped that other organizations supporting institutions involved in the information supply chain will apply to become ISNI Registration Agencies, and would then be able to issue identifiers and interact directly with the ISNI central registry in real time. Incorporating standardized institutional identifiers into business workflows would greatly improve efficiency and cost savings in the information supply chain. The collaboration between the NISO I² initiative and the ISNI Working Group could prove to be a timely and valuable solution to a long-standing problem for information suppliers.

JODY DERIDDER <jlderidder@ua.edu> is Head of Digital Services at the University of Alabama Library and a member of the NISO I² Working Group.
The process of developing a SUSHI client requires testing against the SUSHI servers where usage data is expected to be harvested.

NISO Recommended Practice on Test Modes for SUSHI Servers Issued for Trial Use

A new NISO recommended practice, *Providing a Test Mode for SUSHI Servers* (NISO RP-13-201x), has been issued for a trial use period ending January 31, 2012. The *Standardized Usage Statistics Harvesting Initiative (SUSHI) Protocol* is a NISO standard (ANSI/NISO Z39.93-2007) that automates the retrieval of COUNTER usage statistics by libraries. The process of developing a SUSHI client requires testing against the SUSHI servers where usage data is expected to be harvested. The new Recommended Practice describes how content providers should provide access to their SUSHI Servers in a test mode so that clients can be set up easier and faster, which is of benefit to both libraries and content providers.

“We have seen a tremendous surge in the adoption of the SUSHI standard, especially since it became a requirement for compliance with Release 3 of the COUNTER Code of Practice,” states Oliver Pesch, Chief Strategist for E-Resource Access and Management Services at EBSCO Information Services and Chair of the NISO SUSHI Servers Working Group. “But many SUSHI client developers have encountered difficulty in accessing content providers’ servers to conduct testing of their client software. These recommendations provide guidelines to content providers on how they can easily provide test areas to prospective users of their SUSHI server without providing live, usually confidential, data or placing undue strains on their production servers.”

“SUSHI is quickly becoming one of NISO’s most popular standards,” explains Nettie Lagace, NISO’s Associate Director for Programs. “Libraries that are using the SUSHI protocol have seen significant time savings in gathering their usage statistics. This Recommended Practice will make it even easier for SUSHI to be adopted by reducing and eliminating development roadblocks.”

All content providers who provide COUNTER usage statistics are encouraged to implement the recommendations during the trial and provide their feedback. The recommended practice and an online comment form are available from the SUSHI Server webpage.


SUSHI Server webpage: [www.niso.org/workrooms/sushi/server/](http://www.niso.org/workrooms/sushi/server/)
NISO Receives Mellon Foundation Grant for E-Book Annotation Sharing Workshops

NISO has been awarded a $48,500 grant from the Andrew W. Mellon Foundation to fund two standards incubation workshops, which it will lead with the Internet Archive, on the topic of E-Book Annotation Sharing and Social Reading. These meetings will be held in conjunction with the Frankfurt Book Fair in Frankfurt, Germany, on October 10th, 2011, and the Books In Browsers Meeting in San Francisco, on October 26th, 2011. The Mellon Foundation grant will pay for the planning, organization, and direct meeting expenses for the two meetings workshops, for which NISO will conduct the majority of the planning, organization, and logistical support.

The two workshops will advance the discussions around the system requirements for annotation sharing—including technical challenges of citation location and systems interoperability—and around the development and implementation of a consensus solution for these issues. The objectives of the meetings are to provide input to a NISO-sponsored working group on scope, goals and any initial work the group undertakes; and the advancement of a syntax specification that will be further vetted by a standards working group for how bookmarks and annotations are located and shared in digital books.

Each meeting will include both invited speakers and breakout discussions, and participants will include technologists involved in the development of systems as well as librarians and scholars who would be the direct beneficiaries of annotation and social reading functionality. One goal of holding two meetings that are geographically diverse is to ensure that a world-wide community has opportunity for input and engagement on this issue, since the application of such a technology would be applied across the globe.

For more information, visit: www.niso.org/topics/ccm/e-book_annotation/

Journal Usage Factors Project Publishes Results, Recommendations, and Next Steps

The Journal Usage Factor (JUF) is a proposed new measurement of journal impact and quality that will complement the journal Impact Factor (IF) from ISI and compensate for some of its weaknesses. Unlike the IF, which is based on citation data, the JUF looks at actual usage of an online journal and can begin collecting and reporting data immediately after publication.

**Phase 1** of the project looked at the usefulness and viability of a JUF. Librarians rated a potential JUF second in importance for acquisition decisions and third in importance for retention and renewal decisions. 62.5% of authors felt the IF was given too much weight in assessing authors’ work and 70% welcomed an additional JUF measure.

**Phase 2** of the project tested the proposed JUF formula using real COUNTER data and included a statistical analysis performed by CIBER. Because the data showed high variance in usage between items, it was recommended that the formula be changed to use the median rather than the mean in calculating JUF. A mean usage factor should include a confidence level to address issues of statistical “noise.” A maximum 24-month window for collecting data is sufficient and shorter windows of 6 or 12 months could be considered for the future.

The project concluded that JUF comparisons need to be done within broad subject domains as the usage trend pattern over time varies by subject. The JUF data did not show a statistical association with citation impact and thus provides very different information. The measure is highly subject to “gaming” especially using software agents. The study showed that additional indicators for journal usage half-life or a reading immediacy index might also be useful.

**Phase 3** of the project will use these conclusions for the next steps, which will include preparation of a draft Code of Practice for the Journal Usage Factor, developing an updated subject journal classification taxonomy, and running a trial with a subset of publishers.

View the full report at: www.uksg.org/usagefactors
ISO Focus+ Magazine Now Available for Free Download

The International Organization for Standardization (ISO) has made their ISO Focus+ magazine, available free download in PDF format from the ISO website. Also available are the entire backfiles of ISO Focus+ (from 2010) and the two predecessor magazines ISO Focus (2004-2009) and ISO Management Systems (2001-2009).

ISO Focus+ highlights standardization initiatives, the technical groups that develop ISO standards, and the people who use them. “This initiative will help to further promote knowledge of the global contribution that ISO international standards can make in providing practical tools for tackling global challenges in the service of the international community,” said ISO Secretary-General Rob Steele.

Particular articles that may interest ISQ readers are:

» Introducing XML: A strategic approach to standards publishing (July 2011)
» A Coded World (April 2011)
» How Strong is your Company’s Brand? An ISO standard can help tell you (November 2010)
» Beyond the Barcode: Next generation libraries (April 2010)
» Standardized Standards? The case of the multiple identifiers (February 2010)
» Facilitating Information and Documentation Systems in Changing Times (February 2010)
» Françoise Pellé: Testing new ideas in informal settings (January 2010)

OpenURL Canonical Citation Profile Approved

The OpenURL Maintenance Agency has approved a Canonical Citation Community Profile and Metadata Format for use with the NISO standard The OpenURL Framework for Context-Sensitive Services (ANSI/NISO Z39.88-2004).

The OpenURL Framework Standard defines an architecture for transporting packages of information over a network with the intent of obtaining context-sensitive services pertaining to the referenced resource. The standard is most commonly used to link full-text information resources and discovery search services. The standard included two example community profiles and assigned the maintenance agency with responsibility for approving additional profiles.

The Canonical Citation Community Profile was submitted for registration and approval by David Ruddy, Director, Scholarly Communications Services at the Cornell University Library, as a method for using the OpenURL framework to link canonical citations—to a work or passage within a work that are independent of any specific published edition or translation of the work—with various information services. Examples of canonical citations are: Homer, Iliad, 1:125-130, or Romans 5:19.

The Canonical Citation profile was an outgrowth of a request made to the Cornell University Library by Eric Rebillard, Professor of Classics and History at Cornell, and General Editor of L’Année philologique, an abstracting and indexing service specializing in scholarship about Classical literature. Professor Rebillard was interested in improving the success rate of OpenURLs generated by L’Année, which typically contain citations to specific passages within works of Classical literature. These types of citations are “canonical” because they do not reference a particular edition, but instead use established conventions for citing a work and passages within it. Existing OpenURL implementations assume the existence of a particular manifestation of a work rather than a reference at the “work” level, which is needed for canonical citations. A new metadata format was necessary to carry a description of a canonical citation within an OpenURL Context Object. A grant from The Andrew W. Mellon Foundation was used to explore the possibilities of using OpenURL to provide system independent linking between citations of Classical literature and an increasing array of available online resources in Classics. The resulting metadata format and community profile were then submitted to the OpenURL Maintenance Agency for approval and registration, which was granted in August 2011 following a public review and comment period.

Eric Rebillard and David Ruddy, along with Adam Chandler (Database Management and Electronic Resources Librarian) are currently implementing a Classical Works Knowledge Base that utilizes the new canonical citation linking and provides an intermediary service between a classical text discovery resource, such as L’Année and a user’s OpenURL link resolver. This OpenURL information model that “chains” links resolvers together to provided enriched services could be applicable to other domains as well.

DOI: 10.3789/isqv23n3.2011.08

OpenURL standard: www.niso.org/standards/z39-88-2004/
Canonical Citation Community Profile and Metadata Format: www.niso.org/apps/group_public/documents.php?wg_abbrev=z3988review
Classical Works Knowledge Base: cwkb.org/
In Development or Revision

Listed below are the NISO working groups that are currently developing new or revised standards, recommended practices, or reports. Refer to the NISO website (www.niso.org/workrooms/) and the Newsline quarterly supplements, Working Group Connection (www.niso.org/publications/newsline/), for updates on the working group activities.

<table>
<thead>
<tr>
<th>WORKING GROUP</th>
<th>STATUS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Institutional Identifiers (IT)</td>
<td>Recommended Practice in development.</td>
</tr>
<tr>
<td>Knowledge Base and Related Tools (KBART) Phase II</td>
<td>Phase II Recommended Practice in development.</td>
</tr>
<tr>
<td>Joint project with UKSG</td>
<td></td>
</tr>
<tr>
<td>Physical Delivery of Library Materials</td>
<td>NISO-RP-12-201x, Physical Delivery of Library Materials Recommended Practice issued for public comment through August 21, 2011.</td>
</tr>
<tr>
<td>Presentation and Identification of E-Journals (PIE-J)</td>
<td>Recommended Practice in development.</td>
</tr>
<tr>
<td>RFID for Library Applications Revision</td>
<td>NISO-RP-6-201x, RFID in U.S. Libraries Finalizing for publication following the public comment period.</td>
</tr>
<tr>
<td>Shared Electronic Resource Understanding (SERU)</td>
<td>NISO-RP-7-201X, Shared Electronic Resource Understanding (SERU) Revision to Recommended Practice in development.</td>
</tr>
<tr>
<td>Supplemental Journal Article Materials</td>
<td>Recommended Practice in Development.</td>
</tr>
<tr>
<td>Joint project with NFAIS</td>
<td></td>
</tr>
</tbody>
</table>

DSFTU stands for Draft Standard for Trial Use.
E-books have existed in the library landscape for over a decade, but it is only in the last few years that their use has grown to finally become the game-changer that many have anticipated for so long. E-book availability, distribution, licensing, discoverability, usage, and current and future access require content providers and libraries to adapt many of their existing processes. Amidst this chaos is a wealth of opportunities for new collaborations and initiatives.

FOR INFORMATION AND TO REGISTER, VISIT:
www.niso.org/news/events/2011/ebooks