Office of the Chief IT Architect
NRFC0017 / BCP0008
Category: Best Community Practice

K. Gorlen
NIH
January 2006

Unique Identifier for People: Best Community Practice v1.0

Status of the Memo

This memo specifies a National Institutes of Health (NIH) architecture best practice for the NIH community and requests discussion and suggestions for improvements. Please refer to the current edition of the NIH *Architecture Standards Process* (NRFC 0001) for the standardization state and status of this memo.

Abstract

This memo documents the implementation of a Unique Identifier for People, the **NIH ID Number**, which provides a convenient, public representation of the identities of individuals who use NIH resources and services. It describes how the NIH Enterprise Directory (NED) manages NIH ID Numbers, and presents recommended practices for NIH business managers, application designers/developers, and security officers.

Table of Contents

Approved 1/11/2006

1 Introduction

The **NIH ID Number**, defined in NRFC0005 [1], is a unique identifier for people, which provides a convenient, public representation of the identities of individuals who use NIH resources and services. It is a 10-digit number, managed much like the more familiar Social Security Number (SSN), except that since its use is limited to NIH, its public disclosure does not raise privacy and identity theft concerns. The NIH ID Number is mainly useful as the primary key for people records.

The NIH ID Number is beneficial because it:

- reduces the need for multiple copies of data related to people,
- reduces data entry,
- simplifies database and application development, and
- improves interoperability by making it easier to join people data across independently developed and maintained databases.

The **NIH Enterprise Directory** (NED) [2] enables application programs and users to easily find information about the people who work at NIH. Mainly, NED contains information that **identifies** a particular individual, such as a person's NIH ID number, name, date of birth, place of birth, Social Security Number (SSN), ID photo, and information to **locate** or contact a person at work or home, such as their e-mail address, postal and delivery addresses, telephone numbers, organizational affiliation and status (Employee, Fellow, Contractor, Guest), and so on.

This memo describes how NED identifies and assigns NIH ID Numbers to people, discourages assigning more than one NIH ID Number to the same person, and voids NIH ID Numbers that have been assigned in error.

2 Background

In 1994, NIH formed an information technology Architectural Management Group (AMG) consisting of representatives from each of its Institutes and Centers (ICs). The AMG's broad goal was to define a uniformly supported, interoperable, IT architecture to enable NIH users to transparently access and use the platforms, processes, and data they need to do their work.

The AMG's *Report on Interoperability at the NIH* [3] issued in May, 1997, recommended that "Unique personal identifiers (not the Social Security Number) must be defined".

The NIH Acting CIO subsequently approved the formation of a small Technical Subcommittee, the AMG TSC [4], to further develop the concept and design of a unique identifier for people and an NIH electronic directory service. The AMG TSC was comprised of technical experts from several ICs, who met regularly from August 1997 through November 1998.

The AMG TSC designed a unique identifier for people to reliably associate with an individual all the related information stored in the electronic directory and various other NIH systems and databases. After considering many alternatives and surveying practices at other organizations, the AMG TSC recommended a 10-digit NIH ID Number with the characteristics described in the *Requirements* section of NRFC0005 [1].

The AMG adopted this recommendation at their quarterly meeting on October 15, 1997.

The NIH ID Number with the AMG-approved characteristics was implemented as part of the NIH Enterprise Directory (NED) [5] and deployed for production use in May 2000.

Currently, NED contains NIH ID Numbers for over 80,000 individuals.

3 Implementation

The NIH Enterprise Directory manages NIH ID Numbers as described in this section.

3.1 Identity Binding and Registration

Assigning a single, unique, persistent NIH ID Number to each individual identity requires:

- A single, central database of assigned NIH ID Numbers and their associated individual identifying information, that is, distinctive information about an individual that never or rarely changes; and,

- A central NIH ID sequence number for generating new, unique NIH ID Numbers.

Identity binding is the process of assigning an NIH ID Number to an individual. **Registration** is the process of activating an individual and their bound NIH ID Number in NED.

For individuals meeting the scope requirements of NRFC0005, authorized individuals (such as NIH Administrative Officers) are the **Registration Authorities** (RAs) responsible for performing identity binding and registration.

To help assure uniqueness, singularity, and persistence, all applications must use the following procedure to perform identity binding and registration:

1. Acquire individual identifying information: legal name, alias names, SSN, date of birth, place of birth, gender, and home address. At least a surname is required.

2. Search the single, central NIH ID Number database for a matching identity using the acquired individual identifying information.

3. If the search finds a single, strong match, register the individual using the previously bound NIH ID Number.

4. If the search finds multiple strong matches or one or more weak matches, defer registration until the RA provides additional information.

5. If the search finds no matches, register the individual using a new NIH ID Number generated from the central NIH ID sequence number.

3.2 Record Linkage

The registration search performed in Step 2 above performs **record linkage**, that is, it determines whether two records identify the same individual. NED uses **probabilistic record linkage** [6] [7], which:

- tolerates missing data and typographical errors, and

- quantifies the likelihood of a match by calculating a number that is related to the probability that the match is correct.

Probabilistic record linkage takes into account factors such as:

- the significance of the attributes that agree; for example, two records more likely to identify the same individual if both records have given names of JOHN than if both records have a gender of MALE;

- the closeness of agreement of the spelling of names [8]; for example, JOHNSON more closely agrees with JOHNSEN than with JENKINS; and,

- the discriminating power of agreement on specific attribute values, which is related to how frequently the value occurs; for example, two records more likely identify as the same individual if both records have surnames of ZYWICKE, than if both records have surnames of SMITH.

The identity binding and registration procedure described in Section 3.1 uses the match probability to decide whether a potential identity match is strong (Step 3), weak (Step 4), or a non-match (Step 5).

3.3 Individual Identifying Information

Collecting individual identifying information requires publication of a System of Records notice in the Federal Register. The NED System of Records announcement [9] authorizes collection of and defines permitted uses for the following individual identifying information:

- Legal Name
- Alias Names
- Date of Birth
- Place of Birth
- Social Security Number (SSN)
- Gender
- Home Address

NED uses this information to perform identity binding, as described in Section 3.1. NED also collects and stores the additional individual identifying information authorized in [9] such as an individual's photograph, home and personal mobile telephone numbers, and personal email address, but does not currently use this additional information for identity binding.

Note: The NED System of Records notice also defines the categories of individuals whose individual identifying information can be collected. Obtaining authorization to collect additional individual identifying information, or expanding the categories of individuals covered by NED requires amending the System of Records notice.

3.4 Voiding NIH ID Numbers

It would be necessary to perform a thorough investigation in order to validate an individual's identity and thus guarantee the properties of the NIH ID Number as described in NRFC0005, particularly the singularity and persistence properties. NIH business policies and practices used by RAs to collect and verify individual identifying information are insufficient in this regard. For example, a Contractor may be registered using only a first and last name. Thus, if a Contractor is registered with JOHN SMITH as the only individual identifying information, and that individual separates from NIH and later returns, the previously assigned NIH ID Number is unlikely to be reassigned. In addition, RAs occasionally make undetectable errors during registration, such as entering the wrong individual's SSN, creating invalid or duplicate NIH ID Numbers.

Once bound, NIH ID Numbers are never deleted from the central identity database; instead, invalid and duplicate NIH ID Numbers are marked VOID. If an NIH ID Number is marked VOID because it is a duplicate, the valid, duplicated NIH ID Number is stored in the record of the VOIDed NIH ID Number, thus linking VOIDed NIH ID Numbers to the valid NIH ID Number of the same individual.

4 Recommendations

1. NIH business policies and practices for RAs should be improved to require additional, better-validated identity information for all covered categories of individuals.

2. NED should repeat the registration search to check for duplicate identities whenever the legal name, alias names, SSN, date of birth, place of birth, or gender is changed. (Currently, only a check for duplicate SSNs is performed when the SSN is changed.)

3. When a legal name is changed, NED should record the previous legal name as an alias name, and include alias names in the registration search.

5 Related Standards and Practices

5.1 Personal Identity Verification of Federal Employees and Contractors

Federal Information Processing Standards Publication 201: *Personal Identity Verification (PIV) of Federal Employees and Contractors* [10] describes the minimum requirements for a Federal personal identity verification system that meets the control and security objectives of Homeland Security Presidential Directive 12, and defines a standard to be used by Federal agencies for authenticating the identity of Federal employees and contractors (including contractor employees) for gaining physical access to Federally-controlled facilities (e.g. card access systems) and logical access to Federally-controlled information systems (e.g. single sign-on systems). Identities bound to NIH ID Numbers must therefore be verified as specified by this standard in order to continue to allow them to be used for these purposes.

6 References

[1] Department of Health and Human Services, National Institutes of Health, "Unique Identifier for People", *NRFC0005*.

[2] *NIH Enterprise Directory Home Page*, http://nedinfo.nih.gov.

[3] Report on Interoperability at the NIH, http://nedinfo.nih.gov/AMG/interop.htm.

[4] NIH Architectural Management Group Technical Subcommittee, http://www.alw.nih.gov/Other_resources/amgtech.

[5] NIH Architectural Management Group Technical Subcommittee *Architecture Review*, http://www.alw.nih.gov/Other_resources/amgtech/docs/arch-review/arch-review6.htm.

[6] Fellegi, I. P., and Sunter, A. B. (1969), "A Theory for Record Linkage", *JASA* 40 1183-1210.

[7] Newcombe, H. B. (1988), *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration, and Business*. Oxford: Oxford University Press.

[8] Winkler, W. E. (1990), "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage", *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 472-477.

[9] Department of Health and Human Services, National Institutes of Health, "New System of Records 09–25–0216, Administration: NIH Electronic Directory, HHS/NIH", *Federal Register*, Vol. 65, No. 73, 20181-20183.

[10] "Personal Identity Verification (PIV) of Federal Employees and Contractors", National Institute of Standards and Technology, Federal Information Processing Standards Publication 201, February 25, 2005.

7 Security Considerations

The NIH ID Number raises the following security considerations:

1. The central database of individual identifying information must be safeguarded from unauthorized read/write access. Compromise risks identity theft and imposter attacks.

2. Association of an account with the NIH ID Number of another individual risks an imposter attack: an attacker who can authenticate using the misidentified account can assume another individual's identity.

8 Contact

To contact the NRFC Editor, send an email message to EnterpriseArchitecture@mail.nih.gov.

9 Changes

Version		Change	Authority	Author of Change
0.0	6/9/2005	Original Document		Keith Gorlen
0.1	11/10/2005	Style changes	Keith Gorlen	Michele France
0.2	11/15/2005	Edit to improve readability; clarify Sections 3.4 and 5.1.	Keith Gorlen	Keith Gorlen
0.3	1/9/2006	Fix broken reference in Section 2. Reword Section 3.1 to allow for RAs other than AOs.	Keith Gorlen	Keith Gorlen
1.0	1/11/2006	Document approved.	NRFC0001	Steve Thornton

10 Author's Address

Keith Gorlen (Contractor)
2335 Jones St.
San Francisco, CA 94133-2209

Phone: 415-563-3083
Email: kg2d@nih.gov, gorlen@comcast.net