Data Integrity: The Path Forward

To judge from the number of profiles for a single investigator, human cloning has already reached an advanced stage in NIH's IMPAC II database. A search turned up 196 profiles for a single investigator (who shall remain nameless).

Data Integrity—the accuracy or correctness of the data in a database—is not a significant issue in most areas of IMPAC II, but in terms of personal data, it is a major issue indeed. A recent review of the data demonstrated that less than one percent of overall data in IMPAC II was inaccurate, compared to an 8 percent error rate in personal data.

Errors in data cause endless problems that are compounded by the power of a relational database to propagate error. Errors waste staff time and cause confusion for researchers. The eRA budget calls for $2 million annually to fix these mistakes. If NIH can work out effective ways to prevent the proliferation of erroneous names, duplicated profiles, and other common mistakes, users will be able to do their jobs more effectively.

One solution to the personal data problem is to make Social Security Numbers (SSN) mandatory for applicants for U.S. Government grants. As a single universal identifier (UI), a Social Security Number attached to each record would go far toward unraveling the personal data tangle. Many researchers view this practice as an unacceptable intrusion into personal privacy. Only a small percentage of NIH grant applications include an SSN, and there is little support for implementing such a provision. Approximately two-thirds of the profiles in IMPAC II have associated SSNs.

Some argue that researchers should be urged to cooperate in order to ensure high-quality, speedy handling of their grant applications. Others note that personal information on academic researchers is already publicly available, and even more is available to the U.S. Government, so there is little point in withholding an SSN. However, these are questions of policy and philosophy. eRA must deal with the practical consequences of the current practices, seeking ways to mitigate its effects on the quality of personal data in IMPAC II. Viewed in this light, Data Integrity in eRA becomes a human/technology interface challenge that is connected to important broader social, business, and technological issues.

How Error Creeps In

Where did these problems in the personal ID area of IMPAC II first arise?

A major source of poor data in IMPAC II was the failure to scrub (correct) data when it was, long ago, first entered into the IMPAC I system. Incorrect personal IDs from misspellings, inaccurate SSNs and degree codes were allowed to spread. Nothing can be done to rectify the old mistakes in the legacy IMPAC system. Clearly, the eRA system will benefit when the legacy system is retired and the difficult, expensive, and confounding need to bridge data from IMPAC I to IMPAC II will also be eliminated.

Another inescapable source of errors is the mobility of many academic researchers. When a researcher changes jobs, universities, residences, marital status, Internet service providers, and so on, he or she needs to provide timely, accurate new data. NIH must identify the researcher as matching an existing record rather a first-time PI, and use the new data to update a single, accurate record.

Overly broad business rules are another source of inaccuracy. When users inputting data confront some change or anomaly, they may prefer to start a new record instead of editing an existing one. Or, they may inappropriately edit an old record because it is easier than starting a new one. That may save a few minutes (although even this is not certain), but it creates knotty problems down the road as duplication and incorrect information spread. So part of any concerted effort to reduce the error rate will be to tighten procedures and educate users on how best to deal with anomalies.

person_id vs. profile

In order to devise ways to combat the introduction of errors, it helps to examine exactly how the person_id and profile system operates in IMPAC II's databases. A person_id functions in two ways: as a role_id it is attached to a record; as a profile_id it is a summarization of all the personal information for a single individual.

Person_id	Profile	Pers_type	Name	SSN	Deg	Grant/Cmte
1111111		Profile	Sue Smith	123456789	MD	
1111112	1111111	Project	Sue Smith	123456789	MD	R01CA … -1
1111113	1111111	Cmte	Sue Smith		MD	Scientific …

The first source of error was introduced from the legacy IMPAC I system. If a new investigator was awarded a grant (for example, Sue Smith had R01CA12345-01 and John Doe replaced her on R01CA12345-02) in many cases the name, but not the associated SSN was changed. This results in the following problem in IMPAC II:

Person_id	Profile	Pers_type	Name	SSN	Deg	Grant/Cmte
2222222		Profile	John Doe	123456789	PHD	
2222223	2222222	Project	John Doe	987654321	PHD	R01CA ..-2
2222224	2222222	Cmte	John Doe	987654321	PHD	Scientific ..

Note that the correct SSN may be on the "role" records (project and committee) and the legacy carryover SSN may be on the profile. If the IRDB (which uses data from the profile) is searched, the name will be correct but the SSN will be wrong. Most of the IMPAC II modules provide profile information. A person searching for John Doe, SSN 987654321 will see John Doe, SSN 123456789. There is now a Social Security Number mismatch and the individual can chose to make a new profile. Alternatively the user may look at associated data and decide that the SSN is incorrect and change the SSN. EITHER of these two options may be the incorrect choice!

Other errors come from inappropriately editing an investigators name instead of choosing a new investigator. For example, Jim Smith has multiple records:

Person_id	Profile	Pers_type	Name	SSN	Deg	Grant/Cmte
3333333		Profile	Jim Smith	555555555	MD	
3456789	3333333	Project	Jim Smith	555555555	MD	R01HD..-1
5869878	3333333	Project	Jim Smith	555555555	MD	R01HD..-2
4687900	3333333	Project	Jim Smith	555555555	MD	R01MH..-3
3569879	3333333	Project	Jim Smith	555555555	MD	R01HD..-3

An IMPAC II user wants to change the name on R01HD..-3 to John Doe (our PHD) and updates the SSN. Instead of choosing another profile the user edits the name Jim Smith to John Doe, changes the SSN. The data now looks like this:

Person_id	Profile	Pers_type	Name	SSN	Deg	Grant/Cmte
3333333		Profile	John Doe	987654321	MD	
3456789	3333333	Project	Jim Smith	555555555	MD	R01HD..-1
5869878	3333333	Project	Jim Smith	555555555	MD	R01HD..-2
4687900	3333333	Project	Jim Smith	555555555	MD	R01MH..-3
3569879	3333333	Project	John Doe	555555555	MD	R01HD..-3

Note that in the IRDB (which works on the profile) the grants R01HD..01 through 03 and the MH grant will all now carry the name on the profile that is John Doe. When R01CA..3 (from our first example) is entered; there is now a profile that looks correct to the user. A new grant is added to the profile, with the appropriate degree code:

Person_id	Profile	Pers_type	Name	SSN	Deg	Grant/Cmte
3333333		Profile	John Doe	987654321	MD	
3456789	3333333	Project	Jim Smith	555555555	MD	R01HD..-1
5869878	3333333	Project	Jim Smith	555555555	MD	R01HD..-2
4687900	3333333	Project	Jim Smith	555555555	MD	R01MH..-3
3569879	3333333	Project	John Doe	555555555	MD	R01HD..-3
6823456	3333333	Project	John Doe	987654321	PHD	R01CA..-3

What does this addition do to the profile? It creates an MD/PHD since the PHD is added to the summary:

Person_id	Profile	Pers_type	Name	SSN	Deg	Grant/Cmte
3333333		Profile	John Doe	987654321	MD, PHD	
3456789	3333333	Project	Jim Smith	555555555	MD	R01HD..-1
5869878	3333333	Project	Jim Smith	555555555	MD	R01HD..-2
4687900	3333333	Project	Jim Smith	555555555	MD	R01MH..-3
3569879	3333333	Project	John Doe	555555555	MD	R01HD..-3
6823456	3333333	Project	John Doe	987654321	PHD	R01CA..-3

In addition to the creation of a new MD/PHD (and possibly other personal errors) and incorrect data being retrieved from the IRDB, Jim Smith no longer has a profile. The next search on his name will yield no profile so a new profile will have to be constructed.

Thinking back to our first view of John Doe it becomes apparent that if the SSN had been corrected and the first grant collapsed under the corrected name, that profile could have been chosen, so at this point we would only have 2 John Doe profiles and no Jim Smith profile.

Alternatively, if correcting the SSN on the first John Doe was wrong (it was a DIFFERENT John Doe), then data would still be assigned to the wrong person and other inconsistencies would occur (such as first time investigator, training, etc.).

Alternatively, if the first John Doe SSN wasn't corrected and another John Doe made at that time, there would now be three John Doe profiles! Depending on whether a SSN change was made at the time of the name edit, John Doe could potentially carry 3 different SSNs – 123456789, 987654321, and 555555555.

And so, the errors and inconsistencies grow.