Assignment and Matching in ISNI: a summary

Refer: ISNI Data Quality Policy

To maximise the quality of data, ISNI assignment is only awarded to records that meet (or exceed) ISNI's defined minimal criteria. The rules are different for batch loaded data and online requests.

Assignment

Stages in determining assignment and updating the database:

- Validation. The data must be in UNICODE UTF-8, the date in ISO format, and coded data must adhere to UNESCO LOCLODE, ISO country code, and various elements defined in ISNI data element values.
- Sparse test. If the record is valid, then the record is checked to see it is passes the sparse test. Sparse is defined in the ISNI Data Quality Policy. Note the record may provisionally pass the sparse test but subsequently fail it if the name is found to be a common name
- Search. The identity's name and name variants are searched against the ISNI database. This can involve several steps and several different queries:
    - The data is normalized according to NACO rules
    - For complex personal names (i.e. not simple surname, forename/s or surname, initial/s) name rotation is employed that results in the generation of multiple virtual name variants and multiple queries
    - For personal names, the forenames are checked against a forename equivalence table and more multiple virtual name variants may be generated, e.g. Smith, William will also generate a search for Smith, Bill

  If there is no record retrieved on the first query or set of queries, then a search will be made for the name in truncated form. Not all hits from a truncated search will be accepted as matching candidates where there is a fuller name form that mismatches with the full name form in the incoming record.

  No matching Candidates Found

- Unique. If there are no records retrieved, then the record passes to the unique name test. Personal names are unique if all name variants (surname plus one initial or given name) are unique. Unique names must include at least one full forename or given name. Organisation names are not unique if they are only initials.
- Rich. If the name is not unique, but there are no matching candidates, it is still possible to be assigned if the data passes the richness test. Rich records contain enough information to be capable of matching with future incoming data. Rich is defined in the ISNI Data Quality Policy.
- Single Source. Very few of the sources have this status that is awarded by the ISNI Quality Team and the ISNI-IA Board. It is awarded where the source has high quality records and a very low duplication rate, its file has special significance in the ISNI database and is unlikely to cause duplicates with other sources. Even if a record of one of these designated sources is neither unique nor rich, it will be assigned.

Matching Candidates Found

- If there are records retrieved in the searches, the incoming record is matched with each record as per below.
    - If there is a confident match, the incoming record is merged and an ISNI is assigned (or was already assigned)
    - If there is a match that does not pass the confidence threshold, then in the case of batch load and entry via the web interface, the record is written to the database with a status of provisional and the possible match is recorded in the record. In the case of AtomPub, the record is not written to the database but the information is returned so that the possible match can be resolved in a re-submitted request.
    - If there is no match, the request passes to the rich and single source test (defined above). If records pass one of these tests, they will be assigned except if the record has a possible match field that needs to be resolved.
    - Otherwise, the record is written to the database as provisional or, in the case of AtomPub, returned so that it can be re-submitted in richer form.

ISNI Evaluation and Matching

The following comparisons are made for each record pair:

- Name
- Birth and death date, productive dates, (for organisations – start and end dates)
- Title and partial title (Noise titles are ignored. These are defined as having more than 100 index occurrences in the title index of the ISNI database. Example “collected works”, “Favourite songs”).
- Contributed to or performed (e.g. journal title, musical work)
- Rare title word (currently fewer than 100 index hits)
- Publisher
- Personal affiliation (e.g. co-author) and linked personal identities (e.g. pseudonym)
- Organisation affiliation
- ISBN and ISSN
- Instrument
- Other identity identifier (e.g. ORCID, VIAF)
- Contributor identifier
- Organisation type
- Organisation location (LOCODE or LOCODE conversion)
- Organisation URL

Each test gives a score. The scores are empirically derived from testing with test record sets against the ISNI database and are therefore not relevant outside the context. The evaluation configuration was initially based on VIAF written specifications but tailored after testing. New tests can be added as they are indicated, for example instrument and rare word have been added since the initial configuration.

For names, all the names and name variants in the incoming record are reviewed and the system either chooses the fullest form or constructs the fullest form from the records. The fullest form is then used for comparison. It may match with a shorter form in another record, provided that on that record it is the fullest form.

A global review is then made that assesses the following factors and may lift or lower the overall score accordingly:

- the commonness of the surname (defined as more than 500 hits in the ISNI NA index)
- mismatch of critical tests – name and date. Most of the above tests will have a positive influence but not a negative one, e.g. no match on publisher does not mean that it is not the same identity.
- the overall level of matching. For example, two records each have more than 10 titles and only one title matches. The evaluation considers that it has less confidence in the matching score and lowers it, often making a possible match instead of a match.

A lot is empirical, dynamic and database dependent such as unique name, common surnames, rare title words and noise titles.

Common causes of matching failure

Listed below are the most common causes of matching failure that may lead to provisional instead of assigned status, or duplicate assignment.

- Lack of common elements for comparison (for example one source may include full dates but no titles)
- Approximate dates are given but not coded as approximate (either flourished, meaning was active on that date or circa, meaning "give or take 5 years")
- Other date errors, such as including the starting date of a journal title to which an author contributed
- Spelling mistakes
- Titles including extraneous information

The complete ISNI Quality infrastructure includes regular sampling, crowd sourcing, error notifications, and special programs to assess synchronisation and date anomalies. The matching algorithms are periodically reviewed.

Version	Date	By	Description
1.0	03-07-2017	Janifer Gatenby	First version
1.1	05-07-2017	JG	Additions after review
1.2	06-07-2017	JG	After AMcE review
1.3	12-05-2020	BLQT	Amendments to uniqueness for persons