Our approach to data linkage

1. Data linkage is the process of bringing together records from different data sources that relate to the same individual, object or event. By applying a set of business rules to compare these records, it is possible to determine whether or not pairs of records relate to the same entity. Those pairs of records that do relate to the same entity are deemed to be 'matched'.

2. Within the Office for Students (OfS), we use a fuzzy deterministic method for our data linking. This means we follow exact business rules to determine whether records match or not, but allow for some mistakes or missing data in the records. The datasets we work with are generally of sufficiently high quality to limit the number of missed (false negative) matches.

Datasets

3. The central datasets we use are the Higher Education Statistics Agency (HESA) student record[1] and Individualised Learner Record (ILR) data. In addition, we use a number of subsidiary datasets that provide us with further records, including the Department for Education's National Pupil Database (NPD) (which provides information on school qualifications, free school meals etc.). We have complete coverage of these three datasets from 2002-03.

4. We also use additional databases that already contain linked records, including:

    - Student Loan Company (SLC) data linked with the HESA student record
    - SLC data linked with data from Pearson about enrolments on BTEC Higher National programmes
    - UCAS individualised records (application and acceptance data which we can compare with attendance in HESA and ILR)
    - Longitudinal Education Outcomes data (LEO) which uses administrative tax and benefits data to provide information on graduate employment and earnings.

Data cleaning and standardisation

5. To ensure our data linking is as effective as possible we 'clean' the data to remove potential ambiguities and inconsistencies. We standardise a number of aspects of the data, including:

    - names (e.g. Rosemary, Rosie, Rosy and Rose all become Rosemary; Robert, Bob and Bobby all become Robert), using a list of names compiled for the purpose

[1] In this note HESA student returns refer to both the HESA Student and the HESA Student Alternative records. See https://www.hesa.ac.uk/collection/c21051 and https://www.hesa.ac.uk/collection/c21054.

- grammar (hyphens, capital letters)
- spelling and transposition errors, using the SPEDIS function[2] from the SAS programming language
- formatting (dates, spaces and abbreviations).[3]

6. To improve the accuracy of our data linking, we also use 'parsing' to divide free-form data fields (e.g. name, date of birth and postcode) into separate components prior to linking. For example, we parse names into first, middle, last and maiden names; dates of birth are parsed into month, day and year of birth; and postcodes are parsed into their two constituent parts. Parsing maximises the amount of information available for linking and enables matching to take place when record pairs do not agree character for character. When combined with additional information, these parsed matches may provide sufficient evidence that the records represent the same person.

Linkage

7. We match records between datasets, and across years within the same dataset, using a set of common identifying characteristics chosen for the purpose – for example first name, surname, date of birth, gender and postcode. We decide which identifying characteristics we will use based on the purpose of the linking, the availability of data (e.g. Pearson data has no postcode information) and how confident we are that they will be able to produce a match.

8. To increase computational speed and efficiency we use blocking strategies to restrict comparison of pairs to those likely to match, while taking care not to omit any potential true matches. The blocking we use in our data linkage includes months of date of birth, first initial of surname, and last digit of the HESA unique student identifier (HUSID).

9. We undertake a number of matching processes starting with those identifying characteristics that are likely to give the most accurate matches – for instance the HESA unique student identifier (HUSID) or the UK provider reference number (UKPRN) – and moving to those that are less likely to produce an accurate match, e.g. sex or first name.

10. We use different combinations of identifying characteristics at each step and multiple business rules to determine whether records are matched. The number of matching processes we undertake depends on the availability of the data and the purpose of the linking, although typically we undertake a minimum of four or five.

11. Table 1 illustrates the combination of different identifying characteristics that we use in the matching processes for linking HESA and ILR student data across years. Using these matching processes, a unique longitudinal identifier is created for each individual who appears at any point in the ILR or HESA record.

[2] The SPEDIS function determines the likelihood of two words matching by measuring how close a word is to another in terms of spelling. This is expressed as the asymmetric spelling distance between the two words. We use this function for names and postcodes.

[3] For more examples of variations in linkage identifiers which need addressing, see https://www.ncbi.nlm.nih.gov/books/NBK253312/table/ch4.t1/.

12. The different matching processes may result in a number of true matches between datasets, and across years within the same dataset, all of which relate to the same individual. These matches are resolved programmatically so that a single, unique personal identifier can be assigned to each individual.

Table 1: Identifying characteristics used in matching processes

Records matched on:	Match process				
	1	2	3	4	5
HUSID[4]	✓	✓	✓		✓
UKPRN[5]	✓				
NUMHUS[6]	✓				
HESAINST[7]	✓	✓			
Sex				✓	✓
Surname		✓	✓	✓	
First name		✓	✓	✓	✓
Second name		✓			✓
Birth date		✓	✓		✓
Postcode		✓		✓	
Restriction on common names		✓			✓
Allowance for spelling error in first name		✓			✓
Allowance for maiden name changes and/or spelling errors in surname		✓			✓
Allowance for slight error in birth date				✓	
First part of postcode		✓			✓
Non-contradictory 2nd (middle) name					✓

13. Annex 1 illustrates the sequence of steps that we use to match HESA and ILR student data records in each of the different match processes.

UHN linking using HESA records

14. The first match process we undertake uses a combination of three fields from the HESA individualised student record – UK provider reference number (UKPRN), HESA unique student identifier (HUSID) and HESA student instance identifier (NUMHUS). Together these fields are known as the UHN and uniquely identify a student on a course (or 'instance of study'). The

[4] HESA unique student identifier

[5] UK provider reference number

[6] HESA student instance identifier

[7] HESA institutional identifier

UHN forms a year-on-year linking mechanism which can be used to track the student's progression on the course from one year to the next, from the commencement of study through to completion.

Missing records

15. Missing records within datasets can cause problems in data linkage. Some data linking approaches deal with missing data by either removing these records from the dataset prior to matching or by imputing data (replacing missing data with substituted values).

16. The datasets that we use are generally of very high quality, with little missing data in the 'core' matching variables: date of birth, names, postcode and sex. In most cases we do not remove records from the datasets before matching. If data is missing from the record linking field, matching is not achieved, thereby resulting in a 'missed match' (see paragraph 21). By using multiple matching processes, we significantly improve the effectiveness of our data linkage as we do not rely on one particular field to achieve a match.

17. There are some occasions, however, when we remove records during a particular matching step because they have no chance of matching in that specific step, for example because there is no UCAS number or middle name. Often these fields have a large proportion of data missing.

18. We do not impute data in order to create a data record that we can use in data linking.

Linkage error

19. Table 2 shows all of the possible outcomes that can occur during data linkage.

Table 2: True match status by algorithm output

		Algorithm output	
		Match	**Non-match**
True match status	**Match**	True matches correctly classified as matches	True matches incorrectly classified as non-matches (missed matches/false negatives)
	Non-match	True non-matches incorrectly classified as matches (false matches/false positives)	True non-matches correctly classified as non-matches

20. Two types of error can occur during linking:

- Missed matches (also called false negatives) where records that should match fail to link due to data input error or missing data

- False matches (also called false positives) where records link incorrectly due to different entities possessing similar identifying characteristics (e.g. same date of birth, postcode, sex, etc.).

21. Errors in data linkage occur when the identifying characteristics used in the matching are inadequate to differentiate between records, or when records are prone to missing data, data input errors or changes over time.

22. The likelihood of errors occurring within the datasets that we use is relatively low, as data quality is generally high; those errors that do occur are more likely to be missed matches, for example because of names changing.

23. There are some records for which it is particularly difficult to find sufficient evidence in order to determine a match, for example those people with common names – John Smith, David Jones, Susan Smith, etc. There can also be difficulties with matching records for those people whose self-declared ethnicity is Chinese, as it is traditional for ethnically Chinese surnames to be listed before the forename.

24. Some approaches to data linking remove matched records from datasets to reduce the likelihood of false matches occurring. This is not something we practise (apart from in SLC-Pearson data linking where we are looking for, at most, one match for each individual) as we have no way of knowing whether someone will appear multiple times, or not at all, in the following year's record.

Evaluating data linkage quality

25. It is important to assess and understand the quality of data linkage – any potential errors originating from the quality of the data linkage need to be taken into account when using the linked data for analysis.

26. There are a number of ways in which the quality of data linkage can be assessed:

    - Manual clerical assessment using a sampling approach of record pairs produced by each matching process to assess the quality or accuracy of the link status assigned to record pairs, and assessment of the characteristics of unlinked records
    - Comparison with 'gold-standard' (or reference) data – although this data does not always exist
    - Comparison to previous data matching
    - Comparison to similar data matching produced by other organisations
    - Use of synthetic data that have similar characteristics to the real data, e.g. using purpose-built data generation programmes
    - Systematically corrupting data to generate a dataset for which the ground truth linkage is known, against which the false match rate can be validated
    - Switching off structured matching keys/references in datasets to assess the quality of the data matching using the 'raw' data.

27. Evaluation of the quality of data linkage can be used as an end-stage in the data linkage process to check the accuracy of matches, or it can be used in the ongoing development and refinement of data linkage processes.

28. We do not assess the quality of our record matching as a final step in our data linkage processes; instead, we assess the quality of new data linking methods or algorithms as part of their development. We do this by using manual clerical assessment to study a sample of marginal cases in further detail. As a result, we are able to refine the matching further and make adjustments to our algorithms as appropriate.

29. We routinely use manual clerical assessment to check the robustness of our data linking on an annual basis. We identify where there are large number of records 'clumping' together and undertake a clerical assessment of these particular records to check the data linking. We might also choose to check records where particular fields have changed, for example those records where sex has changed.

30. Every few years we re-run a sample of our data linking using synthetic data – for example, using a dataset in which we have changed dates of birth. By doing this we can check to see whether our original data linking works, and what the level of error is.

31. We sometimes use triangulation with a third party linked dataset (e.g. from HESA or Department for Education) to compare the data linking in these datasets to our own linking. In this way we can identify where there are differences in our linking and can verify the quality of our data linkages.

What does the OfS use data linking for?

32. These are some examples of the main ways we use data linking.

33. **Continuation measures** – we link HESA and ILR student data across years in our calculation of continuation measures. We use combinations of first name(s), surname, date of birth, sex and (where available) home postcode and prior educational establishment. We link each record in the base year to every record we can find for that student in each year's data, to see if the student is still studying at the same provider or a different provider.

34. **Tracking underrepresentation by area (TUNDRA)**[8] – we use the National Pupil Database (NPD) to link pupils from state schools in England to HESA and ILR records to determine the proportion of pupils from each area who later entered university or college.

35. **Equality and diversity**[9] – we use some pupil characteristics from the NPD to identify groups of students who were less likely to enter higher education, or less likely to continue in their course when they started. For instance, pupils eligible for free school meals, which serves as a common proxy for financial disadvantage.

36. **Providers' data returns** – we use data linking to check survey data returns using data from other sources including the Student Loans Company and Pearson Education Limited.

37. **HESES re-creation from HESA data** – we link HESA student data to previous years' data (using the UKPRN, HUSID and NUMHUS triple (UHN)) to help account for definitional

[8] See www.officeforstudents.org.uk/data-and-analysis/young-participation-by-area/about-tundra/.

[9] See www.officeforstudents.org.uk/data-and-analysis/equality-diversity-and-student-characteristics-data/.

differences between HESA and HESES data. The data from earlier years is used to help inform the calculation of FTE for some students.

Future developments

38. We are currently exploring alternative linking methods, in particular 'Splink',[10] an open source probabilistic linking algorithm written in PySpark by the Ministry of Justice. Rather than following a deterministic set of business rules, probabilistic algorithms calculate match probabilities for each pair of records based on their similarity across a number of fields. Pairs of records with a high enough match probability are considered to be matches and others to be non-matches.

39. This change is driven by making our linking more platform-independent, but it may also provide benefits in linking accuracy and will make the approach more transparent to others. This and any other potential new methods will be compared against our current approach to understand the implications on match quality.

[10] See https://www.gov.uk/government/publications/joined-up-data-in-government-the-future-of-data-linking-methods/splink-mojs-open-source-library-for-probabilistic-record-linkage-at-scale#introducing-splink.

Annex 1: Summary of our data matching processes

This annex illustrates the sequence of steps that we use to match HESA and ILR student data records in each of the different match processes.

As described in the main text of this document, in most cases we do not remove records with missing values from the datasets before matching. If data is missing from the record linking field, matching is not achieved, thereby resulting in a 'missed match'. We do not impute data in order to create a data record that we can use in data linking.

By using multiple matching processes (as described in this annex), we significantly improve the effectiveness of our data linkage as we do not rely on one particular field to achieve a match. If a pair of records matches, using any of the matching processes, they are considered a match.

To increase computational speed and efficiency, we use blocking strategies to restrict comparison of pairs to those likely to match, while taking care not to omit any potential true matches.

In this annex, a field listed on its own refers to an exact match on that field. A field listed followed by ≤ X refers to a spelling distance of X or less between the two records. The boxes with 1A, 1B and so on refer to the specific matching process.

Match process one

Data: blocked by HESAINST[11]

Step 1 – records match on institution and student reference number

HUSID[9] and HESAINST/UKPRN[10]

AND

Step 2 – records match on <u>one</u> of the following:

Surname	1A

OR

1st name	1B

OR

Birth date	1C

[11] HESAINST = HESA institutional identifier

Match process two

Data: blocked by birth month; maiden and surname interchangeable

Step 1 – records match on one of the following:

Birth date, HUSID, 1st name ≤ 40, surname ≤ 40

OR

Birth date, postcode, 1st name ≤ 40, surname ≤ 40

OR

Birth date, 1st name, surname

AND

Step 2 – records match on one of the following:

One of: HUSID Postcode First part of postcode (area) Close postcode (spelling distance)	**AND**	One of: 1st name, 2nd name 1st name, surname ≤ 30 1st name ≤ 50, surname ≤ 20 2nd name, surname ≤ 20 1st name ≤ 75, missing 2nd name

2A

OR

1st name, 2nd name, surname — 2B

OR

1st name, surname, and either uncommon name and HESAINST, or very uncommon name — 2C

Match process three

Data: blocked by last digit of HUSID

Records do not match on institution but match on one of the following:

HUSID and surname	3A

OR

HUSID and first name	3B

OR

HUSID and birth date	3C

Match process four

Data: blocked by 1st initial of surname

Records match on:

Sex, 1st name, surname and postcode

And one of the following:

Missing birth date in at least one of the records	4A

OR

Match on birth date	4B

OR

Slight digit error in birth date	4C

Match process five

Data: blocked by birth month

Step 1 – records match on:

Birth date, sex, and 1st part of postcode

AND

Step 2 – for those records that match on postcode, records match on one of the following:

1st name, surname ≤ 5, non-contradictory 2nd name	5A

OR

HUSID	5B

OR

Sex = female, 1st name ≤ 5, non-contradictory 2nd name	5C

OR

Surname ≤ 5, 1st name ≤ 5	5D

Alternative Step 2 – for those records that match on 1st part of postcode, records match on one of the following:

1st name ≤ 5, surname ≤ 5, 2nd name	5E

OR

1st name ≤ 5, surname ≤ 5, Blank 2nd name, uncommon name	5F

OR

Sex = female, 1st name, 2nd name	5G

OR

Sex = female, 1st name, surname ≤ 30	5H