**Methodology for Geospatial Matching**

June 12, 2024

Historic England: Matthew Bristow, Simon Crutchley

Congruence Engine: Daniel Belteki, Nayomi Kasthuriarachchi

* Participants at the meeting developed an agreed strategy for disambiguating mill sites that feature across the 3 HE datasets.
* The agreed-upon method is a combination of a spatial search for sites followed by simple string matching that checks for variations in spelling and reconciles the grid references for sites to 6 figures of both latitude and longitude.
* The spatial search functions as a process of elimination and works as follows:

1. Using GIS software, draw buffers of a small circumference (distance yet to be determined) around the grid references corresponding to the HE **listed** buildings. The circumferences should be large enough so that they capture another point geolocated within the same building complex but not so large that they spill over onto the footprints of surrounding buildings. HE listed building grid references correspond to extant structures (obviously as they are listed) and are immovable. Let’s call this set of buffers the Listing Buffers (LB).
2. Overlay points from the HER dataset (HER) onto LB. Any points from HER that fall within the circumferences of LB points are therefore buildings from HER that contain a listed structure and can be considered the same site. Let’s call this matched up dataset HER\_LB
3. Now overlay points from the Gazetteer dataset (G) onto LB. Any points from G within LB circumferences are buildings from G that contain a listed structure and therefore represent the same site. Let’s call this dataset G\_LB.
4. Subtract HER\_LB from HER to leave a dataset of HER buildings that do not contain a listed asset. Let’s call this HER\_Remain.
5. Subtract G\_LB from Gazetteer to leave a set of Gazetteer buildings that do not contain a listed asset. Let’s call this G\_Remain.
6. Create a buffer around HER\_Remain and overlay G\_Remain. Any matches correspond to Gazetteer buildings without a listing whose location matches up to HER buildings without a listing. Let’s call this HER\_G\_Match.
7. HER\_Remain – HER\_G\_Match = HER points without a listing that also do not correspond to anything in the Gazetteer dataset. Let’s call this HER\_Standalone.
8. G\_Remain – HER\_G\_Match = Gazetteer points that do not contain a listing that also do not match up with anything in the HER dataset. Let’s call this G\_Standalone.

* This brings to an end the spatial search process. Now we move onto text.
* HER\_Standalone and G\_Standalone are likely to be truly separate sites but we can aim to verify this through employing some simple string metrics for checking variations in spelling and through making sure that none of the 6 figure grid references match. This is a simple enough string matching exercise to run via a Python script.
* We discussed the possibility of using HE asset types to filter out all non-mill sites from a dataset in order to expedite the spatial and text matching processes. This sort of expediency may be necessary if we are looking to generalise the process and expand it outside of Bradford but it then raises the question of when the filtering should be done: on the one hand, filtering first would speed up the process but on the other would we lose important historical context that gives rise to interesting research questions by pre-filtering? If the latter is of concern, any filtering should be done after the spatial and text searches have been completed.
* The question of which GIS software to use to construct the buffers and conduct the searches also arose. SC was of the opinion that QGIS appears to lack a lot of the ArcGIS facilities that make buffer construction and subtraction of datasets easy processes. We were wondering whether the spatial searches could be conducted by the UCL Space Syntax team as they may have an ArcGIS licence.
* We should be able to download a full set of text descriptions of assets as part of the dataset that underpins the Historic List Enquiries site. We did try doing this at the meeting but didn’t succeed in the session. The interesting thing about the text descriptions is that the HE personnel writing them denote linked assets using some variation of the abbreviation q.v. e.g. “[Frizingley Hall] Circa 1730-50, a substantial house extending and doubling in size **Frizley Old Hall qv** of similar design”. It should be a fairly easy exercise to extract a list of links through searching for a text string preceding a variation of q.v. MB and SC confirmed that there is often a need to understand the context of the assets linked via q.vs in order to understand the nature of the linkage e.g. the same architect, one building was an extension of another or built on the former site of another etc etc . In order to do this, you need a context-aware search, which could be addressed through using the LLM E-R work developed by CE.