Communication Papers of the 21st Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 37–43
DOI: 10.15439/2026F4286
ISSN 2300-5963 ACSIS, Vol. 49

Potential of Found Establishment Data Linkage for the Social Sciences

Christian Gerhards
0000-0002-8180-3248
Federal Institute for Vocational Education and Training
Friedrich-Ebert Allee 114-116,
53113 Bonn, Germany
Email: gerhards@bibb.de

***Abstract*—Social science research on establishments classically relies on representative survey data from official registers. More recently, digital found data sources that can be analysed with AI, especially geolocatable sources such as online job advertisements and OpenStreetMap (OSM), have opened up new opportunities for establishment research. However, their representativeness and possible biases are often unclear. This paper discusses the question of how the representativeness of operational data from different geospatial sources can be compared. It proposes linking these sources and discusses the practical, methodological, and data protection challenges involved, including AI-supported matching at the individual establishment level. The paper also discusses how generative AI can be used to derive key variables, such as industry classifications, from available information, enabling comparisons with register-based benchmark data. The approach is illustrated by showing how linked geolocatable establishment data from two german projects could be used: OSM-based establishment information and online job advertisement data.**

***Index Terms*— establishment research, business registers, online job advertisements, OpenStreetMap (OSM), AI-supported data linkage.**

I. Introduction: Job advertisement Data and OSM as Data Sources for Establishment Research

INFORMATION at the establishment level, as a meso level of analysis, is highly important for the social sciences, especially in the field of organisational research, labour and education research. The increasing shift of societal processes into the virtual world is making the potential of found data for social-scientific analyses of establishments increasingly relevant. These found data are particularly interesting because, although official register data are comprehensive, they provide only standard information for very basic research questions. Moreover, access to official register data and the possibilities for linking them to other data sources are often restricted. The question of how well data sources cover establishments is of both empirical and methodological importance in order to assess whether, and to what extent, the data allow for representative conclusions. Two heterogeneous examples of found-data sources that have become more important in the organizational, labour and education research in recent years are data from online job advertisements and geodata, such as data from OpenStreetMap (OSM). Data from job advertisements [1] provide a source offering insights into recruitment processes in training firms. OSM, in turn, is an openly accessible found-data source that are predominantly entered by volunteers and that captures business locations, points of interest, and address information on a broad scale [2]. Since neither of these two sources was designed as a complete business register, their analytical value depends not only on sample size, but also on how selectively, completely, and systematically they represent different types of establishments. This paper therefore discusses, how both sources could be linked at the establishment level to assess their suitability for analysing firms and to identify coverage errors.

The proposed approach focuses on linking heterogeneous digital resources through fuzzy matching and record linkage [3], and on using web mining, text mining, and generative AI to derive key establishment characteristics such as industry, firm size, and training activity from names, addresses, job advertisements, legal notices, and linked websites. These derived variables could then be compared with benchmark distributions from register-based data in order to assess data quality, reliability, and bias.

Legal and ethical questions need to be clarified when dealing with establishment data for example because, in the event of re-identification, such data could create a competitive disadvantage for the establishments concerned. Important steps therefore include anonymisation, data protection, and the permissibility of linking OSM and job advertisement data, especially if the enriched data are later used for establishment surveys.

The approach is situated at the intersection of AI-based and computational approaches in the social sciences and economics. It shows how two heterogeneous digital found data sources in the field of labour market and qualification research can be integrated into a critically reflected analytical framework.

ChatGPT was used as a translation aid to support the drafting of the English text.

©2026, PTI

II. State of research and Theoretical Perspective

A. Previous Research on the Selectivity and Coverage

To date, there has been little research on the representativeness (Selectivity and Coverage) —and thus the usefulness —of found data on establishments for the social sciences. Thus, it remains unclear, if online job advertisements can be used as a representative sampling frame for subsequent establishment surveys. The approach proposed here is illustrated, first, by means of a project in Germany that collects online job advertisements: Since 2011, the German Federal Institute for Vocational Education and Training (BIBB) has collected such advertisements from a range of online job boards in Germany collected by a private provider. The project is called KASI, short for Kombinierte Ausbildungsstellenanzeigenanalyse und Inserierendenbefragung — Combined Analysis of Apprenticeship Vacancy Advertisements and Survey of Advertisers [1]. As the project shows, job advertisements do not provide a representative picture of all establishments in Germany. Within this data they identified apprenticeship vacancies using a dedicated classification model. Between 2016 and 2024, the number of identified advertisements increased substantially, while the number of new entrants into vocational training remained relatively stable. In addition, there is pronounced regional variation in the relationship between online advertisements and apprenticeship vacancies reported to the Federal Employment Agency. This selectivity is also confirmed at the international level. The Joint Research Centre (JRC), the European Commission's in-house science and knowledge service, notes that online job advertisements are timely and information-rich, but tend to overrepresent higher-skilled profiles and more formalised, socially desirable skill requirements [4]. For this reason, it is recommended to combine online job advertisement data with survey and administrative sources to asses biases.

Research on OSM likewise makes clear that broad coverage should not be mistaken for completeness. In another project of BIBB Gerhards [2] compares OSM data with the official establishment register of the Federal Employment Agency (BA, short for Bundesagentur für Arbeit) and identifies about 825,000 OSM entries, compared with 2,100,000 establishments in the official BA statistics. Under the assumption that both sources refer to identical units, this would correspond to roughly 39 percent coverage. At the same time, the study shows that although the macro-level relationship between OSM and official BA data is correlated, micro-level validity remains uncertain. Earlier research on OSM in Germany also points to spatially and socially uneven completeness. For retail establishments in the German federal states of Baden-Wuerttemberg and Saxony, Brückner et al. report relatively high average completeness levels, but also a considerable range from 42 to 100 per cent [5]. In particular, large cities and regions with lower unemployment were found to be mapped more comprehensively.

B. Theoretical Perspective on Selective Observability and Unit Incongruence

From a theoretical perspective, the biases of both found data sources can be conceptualised as a problem of selective observability. KASI does not observe establishments as such; rather, it captures a selected subsample: establishments that publicly posted an apprenticeship vacancy online (on established job platforms and on the official website of the Federal Employment Agency) during the observation period, provided an email address, and entered the sample through a selection process. And additionally, as a sampling strategy to avoid duplicate data per establishment, only one training occupation per company—typically the first posted advertisement—is included. OSM, likewise, does not include the full population of establishments (in Germany), but only those business-related units that become visible through mapper activity, data imports, or publicly identifiable points of interest [6]. In both cases, inclusion probabilities are therefore not equal or random, but are shaped by the specific mechanisms through which the data are generated. In the case of KASI, these mechanisms are driven by recruitment practices and digitalisation (strategy to present job postings online in job advertisements); in the case of OSM, they are driven by visibility (potentially establishments that attach importance to being findable via OSM), community activity, and platform-specific logics.

A second theoretical issue concerns the incongruence of observational units. In labour market statistics, the reference concept of an establishment refers to a place of work with employees liable to social insurance contributions; establishments must therefore be distinguished from companies, which may consist of multiple establishments. This distinction is crucial for both KASI and OSM, since in both data sources, individual data points may refer to companies rather than specific establishments. An apprenticeship advertisement may refer to a company as a whole or to several establishments of one company simultaneously, while an OSM entry may represent a branch, company headquarters, showroom, club venue, or another type of point of interest. The central source of bias is therefore not limited to undercoverage, but also includes unit error: entries may formally exist in the data without correctly corresponding to the target population of establishments.

III. Legal and Ethical Considerations in Linking and Reusing Establishment Data

For data on establishments, specific legal and ethical issues arise, particularly in relation to data linkage and the potential subsequent enrichment of these data with additional survey data. In principle, establishment data are often less sensitive than personal data. However, this does not mean that they can be linked, enriched, and reused without legal or ethical constraints. A decisive issue is whether the data remain at the level of organisations or whether they can be connected to identifiable natural persons. This distinction is particularly important when combining KASI and OpenStreetMap (OSM)

data and when considering whether web-scraped contact details may later be used for establishment surveys [7].

A. Permissibility of Linking KASI and OSM

By linking data from KASI and OSM, their representativeness can at least be partially assessed. The linkage of KASI and OSM at the individual establishment level is not prohibited as such, but it is only permissible under clearly defined conditions. The first key question is whether personal data are involved at all. The EU's General Data Protection Regulation (GDPR) protects natural persons, not legal entities [8]. Data relating only to firms in the organisational sense—such as company names, legal forms, or generic contact addresses—are therefore often less problematic. The situation changes where sole proprietors (e.g. "Name Inc."), self-employed persons, named contact persons, personalised email addresses (e. g. name@inc.com), or direct phone numbers are involved. In such cases, the data can relate to identifiable individuals, and matching, enrichment, and subsequent contact may constitute forms of personal data processing.

Where personal data are involved, the crucial issues are the necessity of the processing for a genuine research purpose (this may constitute an exception permitting the processing of personal data), and the proportionality of the procedure. In research contexts, this generally means that only those variables should be linked that are actually required for the analytical objective, and that safeguards such as data minimization (collecting and linking only the data that are necessary), early anonymisation (removing or altering identifying information) where possible, or at least pseudonymization (replacing direct identifiers with codes) should be built into the design. The legal assessment therefore depends not simply on the existence of a scientific interest, but on whether the linkage is genuinely necessary and whether the interests of affected persons are adequately protected [9] [10].

The fact that (found) data are publicly accessible does not remove these obligations. Publicly visible contact details or other business-related information may not be a legally unrestricted resource. If information is scraped or otherwise collected from sources without being obtained directly from the data subject, researchers must generally consider transparency obligations, including informing affected persons about the source, purpose, and legal basis of the processing. Public availability therefore does not automatically legitimise large-scale collection, linkage, or reuse.

In addition to data protection law, other legal constraints may also apply. OSM data are comparatively straightforward in this regard, as they are explicitly made available for reuse under the Open Database License, which allows reuse subject to conditions such as attribution (the data source must be properly credited) and, where applicable, share-alike requirements (derived databases may have to be shared under the same or a compatible open license). It might remains unclear, at least initially, whether data on establishments contained in the two data sources are present without the establishments' consent and may subsequently be processed for purposes entirely different from those for which they were originally made available. KASI raises additional issues because the underlying job advertisement corpus originates from a commercial provider. Even where processing within a research project is lawful, this does not automatically imply that raw data or identifiable components may be freely shared or republished. If further contact details are additionally scraped from company websites, researchers must also take into account website terms of service.

B. Use of Scraped Establishment Addresses for Establishment Surveys

For social science analyses, found data offer new potential alongside designed data from purpose-built surveys. One way to extend the analytical potential of found data is to enrich them with designed data. A possible approach is to use the contact information of establishments (within the framework of the legal and ethical limitations outlined above) — in this case obtained from OSM or KASI — to conduct dedicated establishment surveys. In the case of the KASI project, this is already an integral part and objective of the project. If linked or scraped data are to be used for subsequent establishment surveys, it is important to distinguish clearly between postal contact, email contact, and telephone contact.

Initial contact by post may be comparatively more feasible, but should still be subject to institutional data protection and legal review, especially when it is addressed to a legal entity or a generic business address and the communication is clearly framed as scientific research. In such cases, the research purpose may be more straightforward to assess in the context of institutional data protection and legal review, while the interference with personal privacy may be comparatively limited.

Email contact is more sensitive [11]. Unsolicited electronic communication is subject to stricter legal scrutiny (in the EU and Germany), particularly where no prior relationship or consent exists. Unsolicited contact by email remains legally delicate when it targets named individuals or personalised addresses. Telephone contact is the most sensitive channel [12]. Telephone recruitment based on scraped contact information is generally the most problematic option for establishment surveys.

C. Ethical Guidelines

Even where the legal framework does not prohibit data processing, established research ethics guidelines call for a broader assessment. The ethics guidelines of the Association of Internet Researchers (AoIR), a learned society dedicated to the advancement of the transdisciplinary field of Internet studies, stress that public accessibility should not be equated with ethical acceptability. What matters instead are the context in which the data were made available, the expectations of the persons concerned, the degree of identifiability, the possibility of harm, and the question of whether data use may alter the social situation of those concerned. For establishments in particular, the possibility of re-identification and linkage with data that are otherwise unavailable or difficult to

TABLE I.
PROPOSED APPROACH: STEPS FOR LINKING AND EVALUATING BIASES IN FOUND DATA AT THE ESTABLISHMENT LEVEL

1	legal and ethical considerations in linking and reusing establishment data perspective: permissibility, use of scraped contact information, ethical guidelines
2	Data cleaning: standardization of words, removal of duplicates, fuzzy matching (blocking, similarity measures, manual review of borderline cases)
3	entity resolution (matching of datasets): blocking or candidate generation, pairwise similarity features, human-in-the-loop pipeline (manually coded pairs, active learning, global assignment rule)
4	generative AI for Classification (e.g. on Industry Codes, Establishment Size Groups, and Training Activity) as part of a hybrid pipeline: web research on company websites; substantive enrichment of linked establishments (compilation into a standardized dossier; LLM with restricted classification task); simple cases: Deterministic rules and classical Natural Language Processing (NLP); more complex or ambiguous cases: LLM
5	Assessing Coverage Bias: descriptive coverage matrix, multivariate models

obtain may create competitive risks. Correspondingly, the GESIS as a major German research institute guidance on digital behavioral data highlights privacy, data minimisation, the risk of inferring sensitive attributes, and the need to consider the entire data lifecycle from collection to publication [13].

IV. PROPOSED METHODOLOGICAL APPROACH

Table 1 illustrates the pipeline of the proposed approach for linking and evaluating biases in found data at the establishment level.

A. Error Structure, Data Cleaning, and AI for Fuzzy Matching Between KASI and OSM Establishments

A fair evaluation of the representativeness of KASI and OSM requires a distinction between two establishment-level target populations. The first is the full population of establishments with employees, which serves as the general establishment population. The second is the subset of establishments that advertised apprenticeship vacancies online during the observation period. On both levels the official establishment register of the Federal Employment Agency (BA) can serve as a reference for assessing representativeness. However, due to restrictions, matching at the individual establishment level is not possible, and only marginal distributions can be compared (macro-level comparison). Only at this second level can KASI be interpreted as a coverage instrument (to assess how well it covers the target population) as it captures the subpopulation of apprenticeship-providing establishments that recruit digitally.

Both data sources may contain incorrect inclusions (false positive) as well as missing establishments (false negative).

In KASI, such problems arise mainly from advertisements listed at the company rather than establishment level, multi-location postings, central recruiting mailboxes, or advertisements that cannot be assigned to a specific establishment. Missing cases include establishments that did not advertise online (first level) or have no/could not be linked to a contactable email address.

In OSM, problematic entries may include outdated nodes (individual geographic points), non-economic institutions, brand-only entries, headquarters without a corresponding employee-bearing establishment, or units without employees subject to social insurance contributions. Gerhards [2] notes that the macro-level comparison with Federal Employment Agency data does not resolve whether OSM entries and agency establishments actually refer to the same units, or whether some recorded entries are not employee-bearing establishments at all. In OSM, missing cases include establishments that are unmapped, less visible, or primarily oriented toward business-to-business markets.

Data cleaning must therefore be treated as a separate methodological step. Names, legal-form suffixes, punctuation, umlauts, abbreviations, and address elements first need to be standardized. Duplicates must then be removed within each source, since both repeated job postings and repeated points of interest may occur. Only in a third step should multi-stage fuzzy matching be applied: first, exclusion (blocking) of cases from matching by postal code, municipality, federal state, or spatial distance (in many cases geocoded GPS coordinates are available for OSM and KASI Data); second, similarity measures for names and domains; and third, probabilistic linkage with manual review of borderline cases. External websites and legal notices are especially useful for distinguishing between corporate groups, brands, branches, and local establishments. A validated hand-coded sample is advisable for calibrating thresholds and making precision and recall transparent.

The most obvious and methodologically central application of classical AI (in this step not generative AI) in the proposed approach for linking KASI and OSM is entity resolution [14], that is, the probabilistic linkage of identical establishments across two data sources. OSM contains spatially referenced entries with names, address components, and sometimes websites or phone numbers. KASI contains establishments that advertised apprenticeship vacancies online between July 2024 and July 2025, including contact details and further information extracted from the advertisements. A simple exact match based on names or addresses would fail systematically, because one must expect name variants, branch labels, legal-form suffixes, misspellings, company-level rather than es-

tablishment-level designations, and incomplete location information. For this reason, AI-assisted matching at the individual establishment level represents a key methodological advance.

A multi-stage approach seems well suited for this task and is proposed here. In the first step, strict blocking (exclusion from comparison) or candidate generation (probable matches) should be used so that not every KASI establishment is compared with every OSM entry. Appropriate blocks can be defined by postal code, municipality, district, or a geographic radius around geocoded addresses. In the second step, pairwise similarity features are constructed, including Jaro–Winkler or Levenshtein similarities for names, token-based similarities for street and locality, exact or partial matches for house number, website domain, telephone number, and email domain, as well as spatial distance measures. Domain-specific features are also useful, such as the detection and standardization of common German or international legal-form suffixes like GmbH, AG, e.K., eG, or gGmbH, and branch markers such as "Center," "Market," "Store," "Branch," or place-name extensions. These features can then be combined in a supervised matching model, for example through gradient boosting, random forests, or simple logistic regression.

In practice, the procedure should be designed as a human-in-the-loop pipeline (with active human oversight). First, a gold standard of manually coded pairs should be created, distinguishing confirmed matches, confirmed non-matches, and ambiguous cases. This serves as the basis for training and iteratively improving the classifier. Active learning (a machine learning approach where the model selectively asks for labels on the most informative data points to improve performance efficiently) is especially useful, since the model can present uncertain or contradictory cases for manual review [15]. After pairwise classification, a global assignment rule is required to prevent multiple OSM entries from being incorrectly assigned to a single KASI establishment, or vice versa. The procedure should be evaluated using classification evaluation metrics (precision, recall, F1 score), and calibrated match probabilities, supplemented by error analyses across industry, establishment size class, and federal state.

B. Generative AI for Constructing Variables on Industry, Establishment Size, and Training Activity.

Potential biases in the data should be examined using a multivariate approach. The aim is to compare the marginal distributions in official Federal Employment Agency (BA) data with those observed in the found data of OSM and KASI. To enable this comparison using marginal distributions along classifications of the BA, the information from KASI and OSM must be classified into classes like industry-codes or firm size levels. BA data contains information on industry, location (e. g. German federal states), and employment counts.

Such an analysis requires that key establishment characteristics be available in a consistent form or be reconstructable. For industry, the official WZ classification (short for Wirtschaftszweig, branch, a German system for categorizing economic activities into standardized industry sectors) provides the reference framework.

For KASI and OSM, where direct classification from a reference source is unavailable, industry coding should proceed in stages: through web research on company websites and legal notices; and through text-based classification using job advertisement titles, OSM tags, and activity descriptions.

Federal state can be identified relatively robustly from geocoded addresses or OSM coordinates.

The second field of application concerns the substantive enrichment of linked establishments. This is where generative AI may be especially useful in the approach proposed here. Many analytically central characteristics are not available in KASI or OSM as clean, standardized variables but as texts. Industry, establishment size, and training activity therefore need to be reconstructed or approximated from heterogeneous traces: establishment names, legal form and address details, apprenticeship advertisement text, legal notices, career pages, "about us" pages, and other linked content on company websites. KASI is particularly informative in this respect, using additional information contained in advertisements and a dedicated classifier for apprenticeship postings [1]. Conversely, the OSM study highlights the potential of linking OSM with company websites, job advertisements, and other digital resources [2].

In this context, generative AI should not be used as an open-ended text tool (that generates unrestricted, free-form text without predefined constraints), but as a structured extraction and classification component. A two-stage procedure is methodologically advisable. First, all available text sources and metadata for an establishment should be compiled into a standardized "dossier" (set of variables): name, legal form, OSM tags, advertisement title, advertisement text, legal notice, self-description on the website, contact page, location list, and, where available, career pages. In the second step, a large language model (LLM) [16] should be given a restrictive extraction task with a fixed output schema, for example in JSON (a data format used to store and exchange structured data, such as model inputs, outputs and labels), including the presumed main industry, a WZ-related industry description, indicators of branch structure, indicators of employee count or size class, evidence of training activity, and finally a confidence score (numerical value indicating how certain the model is about its prediction or extracted information), and an "evidence span" field (a reference to the specific text segment(s) in the source data that support or justify the extracted information). This forces the model to derive variables from explicit textual evidence rather than merely guessing.

For example, for industry, terms such as "bakery," "dental practice," "logistics provider," "metalworking," "car dealership," or "tax consultancy" are strong signals. For establishment size, relevant indicators include references to multiple branches, franchise structures, team size, or expressions such as "employees," "locations," "family-run craft business," or "international corporation." For training activity, terms such

as “apprenticeship,” “trainee,” “dual training,” “training occupations,” chambers such as “IHK,” (short for Industrie- und Handelskammer — Chambers of Industry and Commerce) “HWK,” (short for Handwerkskammer — Chambers of Crafts) or “we provide training” are directly informative.

Generative AI should not operate in isolation, but as part of a hybrid pipeline: Deterministic rules and classical Natural Language Processing (NLP) components should first handle the straightforward cases, such as legal-form recognition, domain extraction, pattern matching for training activity, counting location lists, and detecting explicit employee ranges or size information. Only more complex or ambiguous cases should be passed to the LLM, for example when industry must be inferred from several text fragments or when size class must be derived from weaker “signals” (correlations). The procedure may become especially robust when the model does not generate freely, but instead works with retrieval-augmented prompting based only on previously extracted text snippets and is required to return textual evidence for its decision rather than on its internal knowledge alone.

The results should then be filtered using confidence thresholds and validated against external reference sources, such as register-based industry information or known indicators of training activity.

C. Assessing Coverage Bias against a Third “Ground Truth” Source

Once the data have been brought into a comparable form, the analysis for assessing coverage bias should begin with a descriptive coverage matrix. For each source, the first step is to report what share of establishments in the reference data from official Federal Employment Agency (BA) statistics appears overrepresented in KASI, in OSM, in both sources, or in neither. This should be followed by stratified coverage rates by industry, size class or federal state. Raw proportions and comparison of overcoverage and undercoverage makes it possible to show whether a source includes many establishments overall but systematically captures the wrong types of units.

The next step should be a multivariate model, such as a logistic (e.g. training/non training establishments) or multinomial regression (e.g. economic sector). For KASI, the model should explicitly refer to the subpopulation of potentially training-active establishments. For OSM, the general establishment population can also be analyzed.

Three expectations follow from the design. First, KASI is likely to cover the general establishment population much more selectively than OSM, because by design it is restricted to apprenticeship-providing establishments that recruit online and provide an email address. Second, OSM is expected to reach a broader range of establishments, but also to contain more incorrect inclusions, since visible POIs do not necessarily correspond to employee-bearing establishments. Third, the overlap between the two sources is likely to be greatest among larger, publicly visible, service-oriented, or customer-facing establishments, whereas small, industrial, business-to-business, or informally recruiting establishments are likely to remain underrepresented in at least one source. For KASI, findings on regionally uneven online recruitment patterns point to systematic coverage differences. For OSM, earlier research suggests that visibility, urbanity, and local socioeconomic context may matter [2].

V. Conclusion: Linking KASI and OSM with AI: Methodological Insights, Data Quality, and Future Directions

This article suggested an approach how in labor market research found establishment data sources can be linked and compared. It focused how generative AI methods could be used to construct variables from text. The proposed methodological advantage is that generative AI not only can accelerate classification, but also can bridge the gap between unstructured web and advertisement texts and an analyzable multivariate establishment dataset.

Linking found establishment data from online job advertisements (project KASI) and georeferenced data from open street map (OSM) at the individual establishment level promises a double analytical gain: Substantively, it makes it possible to identify which types of establishments are visible in both sources and which are systematically missing. Methodologically, it shows that coverage is not merely a question of quantity, but a question of the unit of observation and the process through which data are generated.

Especially when combined with additional data (e. g. an establishment survey) KASI is likely to be especially valuable for analyzing apprenticeship-providing establishments that recruit online, but not as a general establishment register. OSM, by contrast, is attractive as an open infrastructure for establishment-related found-data research, yet it requires much stricter cleaning and validation before it can serve as a reliable sampling frame.

In this way, the proposed approach would extend the empirical literature in two respects: first, by moving from aggregated coverage comparisons against official statistics to establishment-level error measurement; and second, by systematically combining job advertisement data with open accessible OSM based found data so that the contents of both data sources can be linked.

For the research question on how geodata on establishments can be linked, the linkage of KASI and OSM should be understood not only as a technical task, but also as a legally and ethically conditioned research practice. Establishment data may often be less sensitive than personal data, but they are not free of regulatory and ethical obligations some of which are particularly specific to establishment data. A responsible design therefore requires a careful distinction between organisational and personal information, clear limitation to what is necessary for the research objective, appropriate safeguards for processing and storage, and a cautious approach to any later use of scraped contact details for establishment surveys. For the proposed approach, this leads to a practical conclusion and data protection implications: linking

KASI and OSM is feasible in principle, provided that a clear distinction is maintained between non-personal establishment data and personal contact data or sole proprietorship data, that a sound legal basis is documented, and that research purpose, data minimization, transparency, and the right to object are properly implemented. For subsequent establishment surveys, contacting firms by post or via generic functional addresses is legally and ethically far more defensible than unsolicited phone calls or personalized cold emails. Where additional web scraping of company websites is envisaged, terms of service, technical protection measures, information requirements, a data protection impact assessment for larger-scale processing, and approval by data protection officers or ethics committees should be integrated into the study design.

The main limitation of the design is that two imperfect data sources cannot be fully validated against one another without an external reference. The establishment register of the Federal Employment Agency is the most convincing third source and reference. Another limitation is the focus on Germany-specific data from proprietary or project-specific datasets (OSM and KASI) which may limit the generalisability of the findings to other countries or contexts. Nevertheless, in many countries, OSM data on establishments and online job advertisement data are publicly available or available by web-scraping, making the approach potentially transferable beyond the German context.

A next step in the practical implementation of the proposed approach lies in semi-automated web enrichment, active-learning-based matching procedures, and validation against Federal Employment Agency (BA) reference data. It should be noted: Even the reference source of the BA official register, however, is not perfect. It does not cover solo self-employed persons without relevant employment notifications, and its definition of an establishment may differ from the meaning of an OSM point of interest (POI)/node or a KASI advertisement.

REFERENCES

[1] Bundesinstitut für Berufsbildung (BIBB), “Kombinierte Ausbildungsstellenanzeigenanalyse und Inserierendenbefragung (KASI),” (engl. Federal Institute for Vocational Education and Training (BIBB), “Combined Analysis of Apprenticeship Vacancy Advertisements and Survey of Advertisers (KASI),”) BIBB project documentation, 2025. https://www.bibb.de/de/210741.php

[2] C. Gerhards, “Coverage of OpenStreetMap data at firm level. Assessing the suitability for social science representative firm samples and found data firm analyses,” in Proc. 20th Conf. Computer Science and Intelligence Systems, vol. 43, 2025, pp. 653–658.

[3] Christen, Peter. "The data matching process." Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012. 23-35.

[4] M. Sostero, E. Fernández-Macías, F. D. Baciocco, and E. Cantalini, “The Professional Lens: What Online Job Advertisements Can Say about Occupational Task Profiles,” JRC Working Papers Series on Labour, Education and Technology, no. 2021/02, European Commission, Joint Research Centre, Seville, 2021.

[5] J. Brückner, M. Schott, H. Fan, and A. Zipf, “Assessing shop completeness in OpenStreetMap for two federal states in Germany,” AGILE: GIScience Series, vol. 2, 2021, Art. no. 20.

[6] H. Senaratne, A. Mobasheri, L. Ali, C. Capineri, and M. Haklay, “A review of volunteered geographic information quality assessment methods,” International Journal of Geographical Information Science, vol. 31, no. 1, pp. 139–167, 2017.

[7] M. A. Brown, A. Gruen, G. Maldoff, S. Messing, Z. Sanderson, and M. Zimmer, “Web scraping for research: Legal, ethical, institutional, and scientific considerations,” Big Data & Society, vol. 12, no. 1, 2025, doi: 10.1177/20539517251381686

[8] European Union, “Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation),” Off. J. Eur. Union, L 119, May 4, 2016.

[9] Federal Republic of Germany, “Federal Data Protection Act (BDSG), Section 27: Data processing for purposes of scientific or historical research and for statistical purposes,” official English version, 2025.

[10] European Data Protection Supervisor, “Study on the appropriate safeguards under Article 89(1) GDPR for the processing of personal data for scientific research,” EDPS, Brussels, Jan. 2022.

[11] BVM Berufsverband Deutscher Markt- und Sozialforscher e.V., ADM Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute e.V., and DGOF Deutsche Gesellschaft für Online-Forschung e.V. (engl. BVM Professional Association of German Market and Social Researchers, ADM Working Group of German Market and Social Research Institutes, and DGOF German Society for Online Research.), “Guideline on Online Surveys,” Berlin, 2021.

[12] Rat der Deutschen Markt- und Sozialforschung (engl. Council of German Market and Social Research), “Guideline on Telephone Surveys,” Berlin, 2021.

[13] J. Breuer, S. Stier, J. Lukito, F. Mangold, M. Wieland, D. Radovanović, K. Weller, and C. Wagner, Overview of Ethical Considerations when Working with Digital Behavioral Data. Cologne, Germany: GESIS – Leibniz Institute for the Social Sciences, 2025.

[14] Papadakis, George, et al. "Blocking and filtering techniques for entity resolution: A survey." ACM Computing Surveys (CSUR) 53.2 (2020): 1-42.

[15] A. Primpeli and C. Bizer, “ALMSER: A Graph-Boosted Active Learning Method for Multi-Source Entity Resolution,” in Proc. 20th Int. Semantic Web Conf. (ISWC 2021), 2021.

[16] J. Dagdelen, A. Dunn, S. Lee, N. Walker, S. Rosen, G. Ceder, and A. Jain, “Structured information extraction from scientific text with large language models,” Nature Communications, vol. 15, 2024, Art. no. 1418, doi: 10.1038/s41467-024-45563-x