Person Record Linking for Digital Libraries using Authority Data

Cornelia Hedeler[1]*, Bijan Parsia[1], and Brigitte Mathiak[2]

[1] School of Computer Science, The University of Manchester, Oxford Road, M13 9PL Manchester, UK,
{cornelia.hedeler,bijan.parsia}@manchester.ac.uk
[2] GESIS - Leibniz Institute for the Social Sciences, Unter Sachsenhausen 6-8, 50667 Cologne, Germany
brigitte.mathiak@gesis.org

Abstract. The explicit purpose of Linked Open Data is to link diverse data, or using the web to lower the barriers to linking data currently linked using other methods. Yet, there exist many objects in the Linked Data cloud that refer to the same real world entity, but are not yet explicitly linked. One special case of this are persons, and in particular authors, which may appear in a variety of contexts, but while they often carry many identifiers, the most prominent attempts to link them use auxiliary information, such as co-authors, affiliations, research interests and so on. In this paper, we investigate the possibility to identify the same person in different, previously unconnected digital library and person-centred authority data sets. We use digital library data sets from different domains and authority data sets, test the suitability of auxiliary information for person record linkage and evaluate how difficult it is to re-find the same person.

Keywords: person record linkage, linked data, data quality

1 Introduction and Motivation

Persons are among the most popular object types in Linked data. They appear as authors, actors or simple persons in many popular datasets. However, they are not easy to identify and link to, because the labels given to persons tend to be ambiguous, only cover a fraction of the relevant persons involved or even both. Person names are often unsuitable as identifiers, as more than one person may have the same name (polysemes) or one person may be known under many different names (synonyms), let alone spelling or copying mistakes. Person identifiers are often given in LOD data sets, but their quality is questionable, as person identification even in rich data sets is an open research problem.

Being able to identify two records from two different data sets referring to the same person is useful in a number of cases. First of all, it enriches both data sets.

* Work carried out while author was a visiting researcher at GESIS in Cologne, Germany.

Even if the information is redundant, pointing towards a similar record provides additional value to users. If, however, the information is partially over-lapping, it can be used to gain additional insight, offer more complete data sets and to fuel algorithms that provide even more insight, such as author disambiguation or other ways to improve data quality.

Not all links are of equal value. When the target is an information system, additional information on common queries is desirable. There is often a focus on precision or recall, depending on the purpose of the query. If the data set is providing input for an algorithm, other criteria for data quality apply, such as completeness, error rate and variety of the information. To simplify the inquiry, we value any attached information. Redundant information can be used to increase the quality and trustworthiness of the data, while it also helps with the linking. Additional information is, of course, the real price. We value most the persons of highest interest to the public, although we do not quantify this directly. We instead assume that any subset of persons will be strongly biased towards the most relevant and evaluate those.

In this paper, we compare the linking between a number of data sets. We investigate bibliographic data sets from two domains: Computer Science and Social Science and link to less domain-specific person centred authority data sets: DBpedia and the GND[3] authority file. We investigate how much information is available both on the relevant persons and persons in general and which of that can be used to verify and augment the other data sets. We use that information to perform record linking between the digital library and authority data sets and evaluate this linking.

We find the following:

- The authority data sets contain many person records, but additional structured information is really sparse.
- There seem to be no relevant differences between the domains, regarding the available information (or lack of information).
- Projects such as yago[4] can make a lot of difference, but are not utilised widely.
- Linking works fairly well regardless.

2 Preliminaries and Background

We formulate the person record linkage problem as follows: Let $P = \{p_1, p_2, ..., p_n\}$ be a set of person records in a digital library data set. Each person record has a number of attributes, which amongst others may include publication title(s), co-authors, publisher, venue title, and keywords. For each person record p_i we aim to identify, if present, the person record in the authority data set that refers to the same person.

[3] http://www.dnb.de/EN/lds

[4] http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/

Record linkage generally consists of the following steps: i) data pre-processing to ensure the data to be matched is in the same format, ii) indexing to reduce the number of candidate matches to be compared further, thus reducing the complexity of the matching task, iii) record pair comparison, in which a detailed comparison of the candidate matches is carried out, and iv) classification step, in which candidate matches are classed as matches, non-matches and potential matches, the latter usually requiring manual confirmation or rejection [3]. There is a large body of research on record linkage, and related areas of entity resolution, duplicate record detection and the various aspects of record linkage (see, e.g., [2] for a survey on indexing, [4] for a survey on duplicate record detection, and [11] for a survey on record linkage).

3 Data sets

3.1 Digital library data sets

In contrast to the wealth of metadata available in some digital libraries, the records in the two digital library data sets used here only offer limited metadata, as we will show in the following.

DBLP Most publication records in the DBLP Computer Science Bibliography [6] consist only of author names, publication titles, and venue information, such as names of conferences and journals. In addition to the publication records, DBLP also contains person records, which are created as result of ongoing efforts for author disambiguation [10]. Some of these person records also contain affiliation information, however, that is only the case for 5,936 of the 1,399,292 person records, i.e., for only 0.4% of all person records, making it a fairly limited source of information for person record linkage. In contrast to other digital libraries, such as Pubmed [7], DBLP does not contain abstracts, citation information, keywords or topics. The details of the content of the XML version of DBLP[5] can be found in Table 2.

Sowiport The second digital library data set is a subset of the publication records available on the social science portal Sowiport[6]. Sowiport provides a single source of information on literature and research projects relevant to the social science. For the purpose of this paper, we only focus on a subset of just over 500,000 literature entries in Sowiport that are obtained from three data sources (SOFIS, SOLIS, SSOAR) within GESIS, Leibniz Institute for the Social Sciences, the provider of Sowiport. Publications and projects in these three data sources have been annotated with keywords from TheSoz, a German thesaurus for the Social Sciences. It covers the Social Sciences rather broadly, focusing on sociology, but also containing basic keywords for bordering sciences, such as economics, educational science and political science. It contains 12,000 keywords which are interconnected by several relationships. There are 8,000 'preferred terms', to which less preferred terms are associated as synonyms or almost synonyms with the 'use instead' relationship.

[5] downloaded on 27/04/2014

[6] http://sowiport.gesis.org

Table 1. Number of instances for selected classes

Language	#abstracts	#dbpedia-owl:Person	#dbpedia-owl:Athlete	#dbpedia-owl:Artist	#dbpedia-owl:MusicalArtist
English	4,004,000	831,558	232,082	68,237	37,936
German	1,368,000	119,171	35,824	0	0

3.2 Person-centred authority data

The authority data sets with which the person records from the digital library data sets are to be linked are introduced in the following.

GND authority file and GND publication information As the literature in Sowiport, in particular the subset of the literature selected by us for the purpose of this paper, is heavily biased towards German literature, we decided to utilise the Integrated Authority File (GND) of the German-speaking countries that is offered as part of the linked data service by the German National Library[7] as a linked data dump. With the German National Library being one of the initiators of VIAF, GND is part of the Virtual International Authority File (VIAF)[8]. In addition to the person centred information, the German National Library also provides bibliographic data, which we also use here as additional source of information. Amongst other information, which also includes keywords or subject headings, the GND authority file contains person records, which can either be differentiated (disambiguated), i.e., refer to a real person, and usually contain further information in addition to the name, such as year of birth, profession, or affiliation. In addition to the differentiated person records, there is also a large number of undifferentiated person records, which do not claim to refer to a real person, but are merely placeholders for people with a particular name. Publications linked to a single undifferentiated person record via the author name could in fact be publications by several people who happen to share the same name. As we want to use the GND data for person record linkage, we only use the information that relates to differentiated person records.

DBpedia DBpedia [1, 9, 5], the community effort for extracting structured information from Wikipedia and making it available as linked data, is quickly turning into one of the semantic web hubs for linking data. Table 1 shows the amount and kind of data in version 3.9 of DBpedia[9].

DBpedia is available for download[10] and comes in various data sets containing different kinds of data, amongst them 'Persondata', which contains information about people, such as their date and place of birth and death, and is represented using the FOAF vocabulary, i.e., contains instances of the class foaf:person,

[7] http://www.dnb.de/EN/lds

[8] http://viaf.org

[9] information obtained from `http://wiki.dbpedia.org/Datasets`, `http://wiki.dbpedia.org/Datasets39/DatasetStatistics?v=dqp` and `http://wiki.dbpedia.org/Datasets39/CrossLanguageOverlapStatistics?v=fqa`

[10] http://wiki.dbpedia.org/Downloads39

Table 2. Overview of the content of the data sources used here

Content	DBLP	Sowiport	DBpedia (en)	DBpedia (de)	GND (diff)	GND (undiff)
Person records	1,399,292	(314,375)	1,055,682	479,201	3,004,350	4,664,636
Distinct names	1,398,865	247,054	1,053,504	459,854	2,661,066	4,651,012
Person with affiliation	5936	(108,709)	60,876	1,829	175,763	0
Email address of person	0	0	0	0	0	0
Person with research field/topic	0	0	30,803	0	396,157	0
Person with profession/ occupation	0	0	9,052	178	1,302,501	0
Person with Wikipedia entry	283	0	NA	NA	228,595	0
Person with DBpedia entry	0	0	NA	NA	80,705	0
Person with GND ID	0	0	57,876	115,912	NA	NA
Person with DBLP entry	NA	0	229	37	0	0
Publication/ project records	2,579,264	519,286	NA	NA	3,794,693	3,325,163
Publications with #authors = 1	493,520	344,785	NA	NA	3,288,825	2,937,792
Publications with #authors > 1	2,085,744	174,501	NA	NA	505,868	387,371
Max #authors for a publication	119	42	NA	NA	23	101
Publications with keywords	0	507,283	NA	NA	2,161,544	

rather than of the class owl:person, for which Table 1 contains the statistics. As the persondata subset itself does not contain much additional information, other data sets are required to obtain the information that could be used to aid person record linkage. These data sets contain primarily the information found in the infoboxes on Wikipedia, and is available for download either as the raw infobox data or the cleaned mapping-based data, which means, similarly to GND, DBpedia also has subsets of data of varying quality. The results presented in the following are based on the higher quality mapping-based data.

4 Approach for person record linkage

In the following we provide the details for each of the steps of record linkage. The approach can be seen as preliminary, as the main focus of this work was to evaluate whether there is sufficient information available in such reference data sets to make this a viable approach.

Pre-processing: As both, GND and in particular DBpedia contain information beyond scientific authors, their research interests, co-authors and their publications, pre-processing involves identifying the information relevant for the purpose of author record linkage. The information from all four data sources is standardised, which includes lowercasing of all text, and arranging person names in the

order 'forename middlenames lastname' to achieve a consistent format of the information across the different data sources and ease comparison of the person records for matching.

Indexing: For this step we utilise Apache Solr 4[11], an open source search platform that is built on top of Apache Lucene 4[12], an open source full-text search engine library. To identify potential candidate matches, we index the person names and run a search that takes into account domain specific knowledge. The search accounts for spelling mistakes in either of the given-, (middle-) or lastname, and also allows for missing middle names, the use of nicknames instead of the actual forenames, a swap of forename and last name, and also searches for names where just the initial of the forename is known. All these are variations typically observed in person names. Solr/Lucene uses a mixture of TF-IDF and Levenshtein edit distance as similarity function and is being utilised frequently for information retrieval tasks.

Record pair comparison: The attributes of person records of the digital libraries that are compared with the corresponding attributes of the authority person records are (if present): name of the person, keywords/research interests, affiliations, and names of co-authors.

Classification step: As there is limited information available in the authority person records (see Section 5 for details), the classification step employes a domain specific heuristic to class candidate record pairs as matches, non-matches and potential matches. To exclude spurious person records returned as candidate by the search over the index, candidate record pairs with little similarity in their fullnames (less than a threshold of 0.85 for levenshtein edit distance between the full names of two people) and/or in cases where the information is known those people for whom the year of the publication does not fall within their year of birth (+20 years) and year of death are classed as non-matches. Candidate record pairs with similar full names (greater than 0.85 levenshtein edit distance), with the authority person record marked up as social/computer scientists or that share co-authors, or keywords are classed as matches, and all others as potential matches.

5 Analysis and Evaluation

In the following, we analyse the content and quality, primarily in terms of completeness and whether an appropriate amount of data is available in GND and DBpedia and is valuable for the task in hand.

[11] http://lucene.apache.org/solr/
[12] http://lucene.apache.org/core/

5.1 Analysis of GND

In addition to the (incomplete) list of publications of a person, GND contains additional information that could characterise a person sufficiently. This information includes subject categories that are assigned to the person records, information on their profession or occupation, and keywords assigned to their publications. The subject categories (e.g., 'Personen zu Informatik, Datenverarbeitung'— which could be translated as 'people on computer science, information processing', 'Personen zu Soziologie, Gesellschaft, Arbeit, Sozialgeschichte'— which could be translated as 'people on sociology, society, work, and social history') provide a broad categorisation of the topics a person works in or publishes. Comparing the professions of people who are annotated with a particular subject category, they seem to help aggregate the different professions that publish in computer science, as can be observed in the breadth of the different professions of people with the subject category corresponding to computer science. Occupations include software developer, computer scientist, mathematician, engineer, physician, linguist, but also far less obvious occupations, such as author, artist, historian. However, as can be seen in Table 2 (in the row entitled 'Person with research field/topic') far more people have job descriptions than subject categories, so limiting the pool of potential person records to those with a particular subject category means most likely that some people who would be expected to publish in computer science (or social science), but who do not have one of the obvious job descriptions will be excluded. The same can be observed for social scientists. However, even though more people have their profession assigned to them, as Table 3 shows, there are fairly few people with the professions one would expect to be of relevance when linking person records to authors of publications in computer science and the social sciences. The same applies to the relevant subject categories, especially when compared to the most frequently used subject categories (and professions) that are also shown in Table 3.

5.2 Analysis of DBpedia

As Table 1 already suggested, the German part of DBpedia contains significantly less information, such as mark up with different types of people, that could be useful for author disambiguation. This is confirmed by the data shown in Table 4. The table shows how many out of the 1,055,682 instances of class foaf:person in the English DBpedia and the 479,201 instances of the same class in the German DBpedia are of a type that would suggest that the person might publish and, might therefore be one of the people with publications in either DBLP or Sowiport. As can be seen in that table, hardly any of the German entries are annotated beyond the simple type of class foaf:person, which does not provide sufficient information to include or exclude a person for the purpose of person record linking. The perhaps surprisingly low numbers of instances marked up with more restrictive types that provide more information for the task in hand as a little discouraging and provides first indications that perhaps the information available in DBpedia to date is insufficient for person record

Table 3. Number of person records with selected subject categories and professions of relevance to the context here, and the most frequently used subject categories and professions

Subject category	**# person**
'Personen zu Soziologie, Gesellschaft, Arbeit, Sozialgeschichte'('people on sociology, society, employment and social history')	2,041
'Personen zu Informatik, Datenverarbeitung'('people on computer science, information processing')	277
'Personen zu Malerei, Zeichnung, Grafik'('people on painting, drawing, graphic')	21,386
'Personen zu Literaturgeschichte (Schriftsteller)'('people on history of literature (writer)')	20,546
Profession	
author/ female author	8,319 / 6,301
lecturer / female lecturer	7,931 / 1,204
research associated / female	1,117 / 759
physicist / female physicist	11,595 / 1,280
mathematician / female	7,561 / 908
computer scientist / female	5,443 / 589
sociologists / female sociologist	3,298 / 1,590
social scientist / female	998 / 546
medical doctor	28,192
writer / female writer	26,532 / 11,011

linking and that further work is required to extract more relevant information from Wikipedia.

Unlike the GND authority file and the additional publication records, which are maintained by a library, and therefore, are structured and contain data more akin to digital libraries, making it easier to identify the information relevant in this context, and with it, perhaps making it more naturally fit for purpose, DBpedia was not developed for that purpose. Therefore, the information that is useful for person record linkage is not quite as readily available and needs to be gathered from different kinds of information that might be available about a person. Analysing the properties of foaf:Person and of dbpedia-owl:Person show that foaf:Person has a very limited number of properties and does not really provide any further information beyond the name of a person, and one has to turn to the properties of the dbpedia-owl:Person and the relevant of its subclasses to find further information that could be useful for the task in hand. Table 4 lists primarily properties of the class Person, but also some additional properties, that could provide information suitable for identifying a person for person record linking. We have also included information on the profession of a person, as this might help narrow down the pool of person records to consider. The numbers in the table are those obtained using the cleaned up mapping-based infobox data rather than the raw data. As there is very limited information on publications of people available on Wikipedia, and with that DBpedia, as most people tend to link to their author profiles in digital libraries, or Google Scholar, or their homepage for their list of publications, it is even more vital to have sufficient other

Table 4. Number of instances in persondata for selected classes of relevance to the context here

Class	English	German
# foaf:person	1,055,682	479,201
# dbpedia-owl:person	652,031	215,585
# yago:person	844,562	0
# dbpedia-owl:scientist	15,399	0
# yago:Scientist110560637	44,033	0
# yago:Scholar110557854	137,555	0
# yago:Writer110794014	87883	0
# yago:Intellectual109621545	140,390	0
# yago:Academician109759069	15,074	0
# yago:ComputerScientist109951070	1,667	0
# yago:Mathematician110301261	4,994	0
# yago:Physicist110428004	6,020	0
# yago:SocialScientist110619642	9,083	0
# yago:Philosopher110423589	6,116	0
# dbpedia-owl:Philosopher	1,276	0

information that characterise a person sufficiently. Unfortunately, the numbers in Table 4 confirm the impression of limited availability of information that is important to identify a person unambiguously.

5.3 Evaluation

For a more detailed evaluation of the state of play of the GND and the DBpedia data set and to determine whether the lack of more detailed information described in Sections 5.1 and 5.2 has a negative effect on the performance of person record linking using these reference data, we have taken the following two data set:

- A manually created small test data set for the social sciences consisting of 30 of the top social scientists with a differentiated entry in GND, and with entries in either German or English DBpedia for the majority of them. The GND and DBpedia entries for the social scientists were manually obtained. Only three of the 30 authors do not have an entry in DBpedia, and all of them have an entry in GND.
- A random subset of 250 computer scientists from the set of person records in DBLP that have a link to the corresponding Wikipedia page, from which the corresponding DBpedia entry was derived. The DBpedia entry was then used to obtain the corresponding GND entry (only available for 51 of them) from the person entries in DBpedia with their corresponding GND entries.

and run the person record linkage approach described in Section 4 to identify the GND and DBpedia person record entries for the authors of publications in Sowiport for the first test data set and DBLP for the second test data set. As

Table 5. Number of person instances with selected properties of relevance to the context here

Property	English	German
Author names		
foaf:Name	1,055,682	479,201
rdfs:label	1,055,682	479,201
dbpedia-owl:birthName	44,977	285
dbpedia-owl:pseudonym	1,865	0
Author affiliation		
dbpedia-owl:almaMater	42,318	0
dbpedia-owl:employer	3,232	0
dbpedia-owl:school	1,974	0
dbpedia-owl:university	1,073	0
dbpedia-owl:institution	923	0
dbpedia-owl:college	13,510	1,829
Co-authors		
dbpedia-owl:academicAdvisor	508	0
dbpedia-owl:doctoralAdvisor	3,698	0
dbpedia-owl:doctoralStudent	1,791	0
dbpedia-owl:notableStudent	372	0
dbpedia-owl:influenced	2,830	0
dbpedia-owl:influencedBy	5,928	0
Keywords or topics / research area		
dbpedia-owl:knownFor	17,702	0
dbpedia-owl:notableIdea	392	0
dbpedia-owl:field	17,831	0
dbpedia-owl:significantProject	614	0
Occupation/profession		
dbpedia-owl:profession	9,052	178

there are multiple publications in Sowiport and DBLP for each of the authors, we performed in fact 5,987 and 10,562 matches between an author of a publication and person records in DBpedia or GND for DBLP and Sowiport, respectively, utilising only the information available for that single publication record to compare with the information available for the authority person records.

For the matches and potential matches we calculated the $precision = tp/(tp+fp)$, which is 0.97 for the social scientists with an entry in the German DBpedia, 1 for the social scientists with an entry in the English DBpedia, and 0.92 for the social scientists with an entry in GND. For the computer scientists with entries in DBpedia, we get a precision of 0.89 taking into account the language of the false positives, as we did not check manually whether some of the authors have in fact entries in both English and German DBpedia, and only used the entry that is present in the corresponding DBLP record. For those authors for which we were able to identify their corresponding entry in GND, the precision is lower at 0.7.

Despite the described lack of detailed information in particular in DBpedia, these results of the, admittedly small, evaluation are encouraging and seem to

suggest that the lack of information does not have a too negative effect on the performance of the person record linkage. Some manual inspection of a sample of the false positives suggests a more restrictive threshold on the name of a person, which, however, limits how noisy the data in the digital library can be for the person record linkage still to work. However, the data set used here is fairly small and does not contain too many people with common names, which are those that contribute mainly to the false positives.

To evaluate how much the name of a person and how much of the additional information (if available) on GND and DBpedia contributes to the correct matching of authors to their corresponding person records, we determined for how many of the correctly matched authors of a publication records, the corresponding reference person record has information on co-authors or keywords and for how many an overlap between those in the person record and those in the publication record was detected.

The results are as follows: for the social science test set we found that out of 5226 publication records matched correctly to person record in the German DBpedia and out of the 1646 matched to the English DBpedia 0 benefitted from overlapping topics for the German DBpedia, which was caused by the fact that none of the German DBpedia entries have topics associated with them, and only 8 showed an overlap for the English version, whereby 183 of the English DBpedia entries have topics/keywords. The overlap for the English DBpedia improves slightly to 43 when compared with the classification assigned with publications in Sowiport, which is at a more abstract and less detailed level than the keywords assigned to the publications. None of the matches showed any overlap in the co-authors, as none are available for the person records in either DBpedia version. Out of the 5580 publication records matched correctly to their corresponding person record in GND, all of them have co-author and 5169 have keyword information available in GND. These numbers are so high as both information was obtained from all the publication records assigned to the respective person record in GND. An overlap in the keywords with the more abstract classification in the Sowiport publication records was found for 374 records, and an overlap for the keywords in 3726 records, but for none of them an overlap in co-authors.

As DBLP does not contain keyword information, we are only able to study the contribution the co-author information makes, and observed that out of the 8679 matched correctly to entries in DBpedia and 1985 matched correctly to entries in GND, none had information on co-authors in the German DBpedia or English DBpedia, but all of them had co-author information in GND, with 26 of them showing an overlap. This, unfortunately, confirms the observation that there is currently very limited information beyond the author name that could help characterise a person sufficiently unambiguous for person record linkage.

6 Discussion

The analysis and evaluation of DBpedia and GND has shown that the semantic markup of the information in DBpedia is still lacking in various aspects, for example, there is a lack of detailed information that is sufficient to characterise a person. Most information tends to be at a higher, more abstract level, which contains less useable information content for tasks such as person record linking (e.g., annotation of a person as scientist rather than social scientist or computer scientist). Furthermore, coverage or tuple completeness and population completeness seem to be generally lacking, whereby determining the extent of the latter requires further analysis of the data sets.

How much of an issue this lack of appropriately detailed information and lack of completeness really causes for tasks does not only depend on the corresponding subset of the reference data and its properties, but also on the remainder of the reference data set, and the digital library data set. The remainder of the reference data set plays a role in terms of how many people share their name with people in the subset of interest, and whether there is sufficient detail available for those people in the subset to determine that they are the people that have indeed written certain publications, and/or whether there is sufficient detail available for the people in the remainder of the reference data set to exclude those people with confidence. The latter, however, is the harder part, as the fact that a person has a profession listed that seems unlikely to lead to a particular paper, still does not provide sufficient evidence to rule out a person with high confidence (see also the list of professions of people who publish in computer science or social science listed in Section 5.1).

This would suggest that a quality measure that assesses the suitability of the reference data set for author disambiguation should take into account the following:

- The tuple completeness, i.e., in terms of how many of the different kinds of properties and information that have been identified as sufficient to characterise a person, are available (across the whole data set).
- The specificity of the annotation with ontologies, i.e., whether the annotation is with rather abstract, generic terms, or as it should ideally be the case, with the most specific that is appropriate in that context, as the use of ontologies enables the retrieval of the more generic terms if needed.
- How much of the information is provided in form of ontologies or thesaurus or even worse literal strings, which provides an indication of the expected heterogeneity of the information across different data sets.
- The number of people in the reference data set who share their names.

To bring this into context with the digital library data sets, one could also determine whether and how many of the author names are shared with several person records in the reference data set. In particular in these cases, sufficiently detailed information is vital in order to be able to identify the correct person record or determine that there is no person record available for that particular person, even though there are plenty of records for people with the same name.

7 Conclusion and future work

There are challenges and opportunities with using the authority data sets for person record linkage, like DBpedia and GND. The challenge is that they cover a far wider spread of different people than a single digital library covering generally a single, if rather broad, subject area. This means, that using DBpedia or GND or similar data sources introduces extra noise in terms of all the additional people, who might share their name with someone in the digital library, but who for whatever reason cannot be ruled out with sufficiently high confidence. This is particularly tricky if the actual person does not have an entry in DBpedia or GND, as neither of them are complete, in which case it is not even possible to compare the potential candidates in terms of the information available about them in DBpedia or GND to determine which one of them is more likely to be the correct person record.

For people with very common names, but who are the only person with that name publishing in a particular area covered by a single digital library, external data sources that cover a wider area, like DBpedia and GND might actually make it harder to identify the corresponding person record as they provide additional people with the same name. These data sources may not contain sufficient supporting information to identify the correct person and using seemingly negative information that could provide clues to exclude a person as a candidate might lead to excluding the correct person if not used with care, as the information included in these sites might only present one particular aspect of a (e.g., multidisciplinary) person. With more research becoming more multidisciplinary, this issue might increase, unless the reference data sources are kept sufficiently up-to-date and contain sufficient detail to cover all subject areas a person works in.

In comparison to (semi-) automatically maintained digital libraries, manual efforts, such as GND, will always result in better quality in terms of their accuracy, but will always lag in terms of their completeness. However, community driven efforts such as Wikipedia and with it DBpedia, might provide the necessary information to fill the gap. To improve its usefulness for approaches that harness the information automatically and do not want to rely on linguistic analysis of the Wikipedia texts, more systematic use of ontologies and thesaurus is needed to annotate the information systematically. In addition, systematic mapping efforts between different thesaurus and ontologies are also needed to be able to utilise DBpedia to map between information in different digital libraries.

High confidence mapping between entries in digital libraries and DBpedia or GND can be added to the digital libraries, providing additional information and value to the end-users of the digital libraries. For those entries for which multiple potential candidates in DBpedia or GND are found, these potential candidates could be provided as further information to librarians or other people who maintain the libraries and help with author disambiguation, thereby, providing further information that could potentially help identify the correct candidate or rule out others.

As next steps we plan to analyse the raw infobox data from DBpedia and expand on the use of the GND authority file to include other authority files contributing to VIAF and determine how much additional information these expansions provide, but also how much extra processing of the data is required to make it suitable for person record linkage. Furthermore, we plan to utilise mappings between keywords of different thesaurus (e.g., those provided by the KoMoHe project [8]) and evaluate the benefit these mappings provide for the task in hand, in particular in light of the other digital library data sets that are integrated in Sowiport and not annotated with TheSoz, but with keywords from other thesaurus. We also plan to expand the data sets used for evaluation to gain a more representative indication of the effect the information available or the lack thereof has on the performance of person record linkage.

Acknowledgments. This work is funded by the DFG as part of the project 'Smart Harvesting'. We acknowledge their support.

References

1. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S.: DBpedia-A crystallization point for the Web of Data. "Web Semantics: Science, Services and Agents on the World Wide Web" 7(3), 154–165 (2009)
2. Christen, P.: A Survey of Indexing Techniques for Scalable Record Linkage and Deduplication. IEEE Transactions on Knowledge and Data Engineering 24(9), 1537–1555 (2012)
3. Christen, P.: Data Matching Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Springer (2012)
4. Elmagarmid, A., Ipeirotis, P., Verykios, V.: Duplicate record detection: A survey. IEEE Transactions on Knowledge and Data Engineering 19(1), 1–16 (2007)
5. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S.: Dbpedia-a large-scale, multilingual knowledge base extracted from wikipedia. Semantic Web Journal (2013)
6. Ley, M.: DBLP: some lessons learned. In: Proceedings of the VLDB Endowment. VLDB Endowment (2009)
7. Liu, W., Islamaj Doğan, R., Kim, S., Comeau, D.C., Kim, W., Yeganova, L., Lu, Z., Wilbur, W.J.: Author name disambiguation for PubMed. Journal of the Association for Information Science and Technology (2013)
8. Mayr, P., Petras, V.: Building a Terminology Network for Search. In: International Conference on Dublin Core and Metadata Applications. Humboldt-Universität zu Berlin (2008)
9. Mendes, P.N., Jakob, M., Bizer, C.: DBpedia: A Multilingual Cross-domain Knowledge Base. In: LREC. pp. 1813–1817 (2012)
10. Reuther, P., Walter, B., Ley, M., Weber, A., Klink, S.: Managing the Quality of Person Names in DBLP. In: ECDL 2006. pp. 508–511. Springer Berlin Heidelberg, Berlin, Heidelberg (2006)
11. Winkler, W.E.: Overview of record linkage and current research directions. Tech. Rep. Research Report Statistics #2006-2 (2006)