Linking Geographical Vocabularies To Fulfil Information Seeking Needs In The Cultural Heritage Domain

Daan de Ruijter[1]

Vrije Universiteit Amsterdam
d.a.c.de.ruijter@student.vu.nl[1]

Abstract. Being able to align concepts from different vocabularies to each other is an important requirement for extracting value from linked data. Vocabulary alignment methods have been studied to a point where relatively straightforward alignments can be generated with a high level of confidence. However, cases remain where the alignment methods fail to generate conclusive results. This paper explores the possibilities of aligning geographical vocabularies using various state-of-the-art alignment techniques, with an emphasis on exploring various disambiguation techniques for when too many mappings are returned. Additionally, the alignments made are utilised in the development of a generous interface to identify to which degree generous interfaces combining metadata from different vocabularies can address challenges relating to information seeking needs in the cultural heritage domain.

Keywords: Vocabulary Alignment · Generous Interfaces.

1 Introduction

The manner in which organisations process, store and access data influences their performance and capabilities. One option organisations can choose to provide an underlying structure for large amounts of heterogeneous data is with a structured vocabulary. Examples of structured vocabularies include ontologies, taxonomies and thesauri [4]. Concepts from different vocabularies can be matched to one-another in a process referred to as vocabulary alignment [13]. These alignments can be used to make vocabularies interoperable with each other (e.g., when creating a system using multiple data sources), or to mutually enrich concepts in a knowledge base through linked metadata. These structured vocabularies are integrated in most of the archival work done by institutions from the cultural heritage domain [6]. However, the heterogeneity of the data represented by the vocabularies in this domain make alignments between these vocabularies challenging [13,15,6]. This heterogeneity can also prove challenging when displaying or querying data from the domain.

Cultural heritage experts can be viewed as people working with cultural heritage data for professional reasons (e.g., research). Additionally, cultural heritage

experts can be people working directly for a cultural heritage institute, or working together with such an institute. The type of cultural heritage experts considered for this research are researchers working together with a cultural heritage institute. A common way for these types of cultural heritage experts to obtain the data needed to fulfil their information needs is using interfaces provided by cultural heritage institutions. For example, an information need for experts in the cultural heritage domain is to be able to find interesting connections between metadata [2]. However, traditional search tools provided by cultural heritage institutions typically do not support these complex and high-level information needs [2,1].

The shortcomings of traditional search are not limited to the cultural heritage domain [16,17]. In his work, Whitelaw proposes an alternative way of presenting data to users by the name of "generous interfaces". Instead of only showing results after a user has entered a search query, Whitelaw argues that users should be immediately presented with interesting data, connections between data, and options to further manipulate the presented data. In doing so, users unfamiliar with the data gain the ability to explore it. This approach is gaining traction in the cultural heritage domain, where the online collection is often the only point of access for users [16,18]. An example of a generous interface in the cultural heritage domain is an interactive interface displaying the coins collection of the Münzkabinett Berlin[1]. This high quality example illustrates a collection presented according to the main principles of generous interfaces.

Section 1.1 describes a number of challenges identified in the fields of vocabulary alignment and generous interfaces. The rest of this document presents a number of findings from related studies in Section 2. Research questions addressing the identified challenges are given in Section 3. The methodology used to produce the alignments and its results are presented in Section 4 and Section 5. The setup for building and evaluating the Generous Interface is given in Section 6, the interface and its evaluation are described in Section 7. Section 8 discusses the implications of the findings in a broader sense. Finally, Section 9 briefly summarises the main findings of this research and provides directions for future work.

1.1 Problem Statement

This research provides a contribution to previous work by addressing two identified research challenges. Firstly, methods for aligning geographical concepts are further developed by addressing challenges related to the disambiguation of alignment results. Secondly, guidelines for visualising enriched geographical metadata in a generous interface and the benefits of such visualisations are described.

Geographical disambiguation techniques in vocabulary alignment. Vocabulary alignment is an ongoing topic of research which has been studied in various contexts like the Semantic Web, cultural heritage or databases [10].

[1] `https://uclab.fh-potsdam.de/coins/`

Concepts described by structured vocabularies from the cultural heritage domain have relatively few connections between them (sparsity), and describe a relatively large variation of objects (heterogeneity) compared to vocabularies from other domains [15]. This sparse and heterogeneous nature of vocabularies from the cultural heritage domain offer promising alignment opportunities [3]. For example, while general methods for finding alignments between vocabularies are described in detail [5,10,13,14,15], techniques for the disambiguation of one-to-many alignments are only described on a surface level.

Disambiguation of one-to-many alignment results can be applied when multiple mappings between a concept and an external data source are produced. The goal of these disambiguation techniques is to identify the single result that is the correct alignment for the concept in question. A possible disambiguation technique is, for example, to find similarities between the hierarchy of the concepts and the hierarchy of the concepts that it is mapped to [13].

Intuitively, a disambiguation technique somehow leverages the metadata of the concept and the returned concepts to find similarities between individual concepts. For a disambiguation technique to work consistently, this would thus require the metadata of both the source and the target vocabulary to be complete and accurate. The source vocabulary used in this paper, namely the geographical concepts in the GTAA[2], offers next to no metadata about its geographical concepts. This limits the possibilities of implementing existing disambiguation techniques. For example, the proposed disambiguation technique of comparing concept hierarchies is not viable, given that no hierarchy is defined between the geographical concepts in the GTAA.

Generous interfaces in the cultural heritage domain. Linking data to external sources allows for the enrichment of concepts by extracting additional metadata from the concepts of these external sources. This metadata allows for concepts to be used for additional purposes. Having a hierarchy between concepts for example, allows for more meaningful navigation between concepts or for automatic inferences to be made about the relation between concepts. One such purpose that particularly relies on the existence of rich metadata is for the data to be displayed in a generous interface [16].

Generous interfaces are an increasingly popular method of providing users access to (digitised) cultural heritage collections. However, challenges regarding data requirements for generous interfaces are still a topic of ongoing research [16,2]. For example, generous interfaces generally require more data as well as computation power. As such, traditional query results may not be able to provide the data in an efficient manner. Additionally, the evaluation of generous interfaces in the cultural heritage domain with regards to their ability to address issues regarding information needs for cultural heritage experts is an area that requires further research. While the shortcomings of traditional search tools are well documented [16,2], it is not clear if generous interfaces are the solution to these shortcomings. For example, generous interfaces could be too complex

[2] `http://data.beeldengeluid.nl/api/collections/beng:gtaa.html`

(compared to standard search tools) for cultural heritage experts to make effective use of.

2 Related Work

This section provides an overview of relevant research for this paper on a number of topics. For each topic an overview of important findings is given, as well as an elaboration on how this paper builds upon these findings.

2.1 Vocabulary quality assessment

A number of measurable vocabulary quality issues are defined [8,12]. While these quality issues are based on SKOS[3] vocabularies, the majority is applicable to most structured vocabularies formats (e.g., having empty labels). Analysing a vocabulary on these issues allows for a prediction of its interoperability with other vocabularies. In other words, a vocabulary with a low amount of quality issues can be more easily aligned with other vocabularies compared to a vocabulary with a high amount of quality issues.

Whether or not these quality issues are present in a vocabulary can be assessed using a number of tools like the PoolParty SKOS Quality Checker[4] or Skosify[5]. Both of these tools are based on the quality issues analysed by qSKOS[6]. The vocabularies used in this research are analysed on these quality issues in order to assess their interoperability and to validate the number and quality of alignments made.

2.2 Vocabulary alignment

Alignment methods can leverage different aspects of concepts in a vocabulary [10,11]. For example, terminological matching compares the labels describing objects and scores their similarity according to some string similarity metric. Other aspects such as structural matching compare similarities in hierarchy or relations between concepts.

The general recommendation for producing alignments between vocabularies is to use an iterative loop of evaluation and enhancement of alignment results [5,15]. In line with this recommendation, this paper takes an iterative approach to producing an alignment strategy between vocabularies. A more specific recommendation is to incorporate at least three high-level steps in the alignment strategy [13].

1. Produce baseline alignment results using exact label matching.

[3] https://www.w3.org/TR/skos-primer/
[4] https://qskos.poolparty.biz/login
[5] https://skosify.readthedocs.io/en/latest/
[6] https://github.com/cmader/qSKOS/wiki/Quality-Issues

2. Identify overlapping alignments using more ambiguous lexical and structured techniques compared to exact label matching.
3. On the one-to-many mappings of these overlapping alignments, apply disambiguation techniques to identify the correct one-to-one mapping.

Methods for steps one and two are generally well defined. For example, a number of string metrics have been described for the purpose of terminological matching between concepts [11]. However, the description of disambiguation techniques performed in step three are only described on a surface level. This paper dives deeper into this subject in order to find some additional insights into the effectiveness of disambiguation techniques.

2.3 Information seeking needs in the cultural heritage domain

Sources used by experts in the cultural heritage domain cover a wide variety of media like video, images or plain text [2]. Additionally, these sources often combine media with various degrees of structure. Simple fact-finding search tasks on a single source are supported by most tools in the cultural heritage domain. However, the majority of search tasks performed by these experts are relatively complex and high level. As such, traditional search does not support the complex information needs of cultural heritage experts from these rich and heterogeneous sources [2,1].

While experts in the cultural heritage domain often need to aggregate or compare results from different sources, the majority of tools allow access to only a single source at once [2]. Solutions that allow experts to request results from multiple sources at once exists (e.g., the CLARIAH Media Suite[7]), but these solutions do not address all complex and high level search tasks of the experts [1]. Given that a large portion of search tasks consists of comparing objects and relations between different sources, advances in vocabulary alignment between sources in the cultural heritage domain could help shape a more appropriate solution that addresses these search needs. This paper puts focus on finding an alignment based solution for a use-case related to the information seeking needs of cultural heritage experts.

2.4 Generous interfaces

One promising direction for the creation of solutions for the complex and high level search tasks executed by cultural heritage experts are generous interfaces [16]. Generous interfaces aim to create more exploratory ways for users to interact with data, compared to the standard search applications which require manual user input before any results are returned [1,16].

A key benefit of generous interfaces compared to traditional search solutions is that they offer a browsable overview of metadata of the concepts in a data source [16]. In doing so, generous interfaces allow users to at a glance obtain

[7] `https://mediasuite.clariah.nl/`

an overview of various general aspects of a collection. With this overview, even users without prior experience of working with a collection can easily grasp the size and scope of a collection and identify where they might find objects they are interested in. Additionally, generous interfaces encourage users to perform exploratory search tasks by offering the option to explore evocative examples of objects and relations between them. By focusing on visualising the relations between various objects or concepts, generous interfaces attempt to address the search task related problems faced by experts in the cultural heritage domain.

Based on the shortcomings of traditional search solutions for experts in the cultural heritage domain identified in [2], Amin *et al.* build and tested an interface that focused on comparison search for linked cultural heritage sources. In doing so, this interface shares a similar design philosophy with generous interfaces. The findings concluded that cultural heritage experts perceived such an interface to be easier to use compared to the baseline tool provided by the study. By tackling a specific use-case from this domain, this paper aims to provide additional results that verify the viability of generous interfaces for these type of search tasks.

3 Research Questions

In order to address the research challenges identified in the previous sections, the following research questions are defined:

1. How does the inclusion of geographical based one-to-many disambiguation techniques affect the performance of current state-of-the-art vocabulary alignment methods for aligning geographical locations from separate structured vocabularies?
   (a) What is the performance of an exact string matching baseline alignment?
   (b) What is the impact on performance of sophisticated alignment methods compared to the baseline?
   (c) How do the performances of various one-to-many alignment disambiguation techniques compare to each other and the baseline?
2. To what extent can issues regarding information seeking needs in the cultural heritage domain be addressed by displaying geographical data enriched via alignments in a generous interface?
   (a) Which issues regarding information seeking needs in the cultural heritage domain, as identified by [2,16], can be expected to be addressed by generous interfaces?
   (b) How are solutions offered by generous interfaces to use-cases regarding information seeking needs in the cultural heritage domain evaluated by domain experts?

4 Alignment Method

To answer the first research question, alignments are made between two real world structured vocabularies containing geographical data. The source vocabu-

lary is the geographical concept scheme from the General Thesaurus for Audio-visual Archives (GTAA), created and maintained by The Netherlands Institute for Sound and Vision[8] (NISV). The target vocabulary is the GeoNames geographical database created and maintained by Geonames.org[9].

Initial assessment. Before performing the alignment strategy outlined in Section 2.2, an analysis on the characteristics and possible quality issues of both vocabularies is done. This analysis is performed in order to validate the viability of various alignment strategies, as well as to gain insight into possible limitations of this research. The quality issues are based on the quality issues implemented in qSKOS, which are explained in further detail in the source material [12]. Examples of assessed quality issues are the incompleteness of label coverage, or the existence of inconsistencies in the hierarchy between concepts. The analysis is done using Skosify, as well as custom python scripts for analysis techniques not supported by standard tooling.

Aligning GTAA and GeoNames. After an initial assessment of the structure and quality of both vocabularies, alignments between the two are produced by adapting the general alignment strategy proposed by Tordai *et al.* [13] (introduced in Section 2.2). Firstly, an initial set of alignments is produced by means of exact string matching between the preferred labels of both vocabularies. Secondly, the extent to which this initial baseline can be expanded by means of more ambiguous alignment techniques is explored. These additional alignment techniques can be roughly segmented into two categories:

1. Complete alignment tools, either discussed by [10] or tooling specifically introduced for use in the cultural heritage domain (e.g., Cultuurlink[10]).
2. Custom implementations of techniques belonging to one of the four categories (*lexical*, *structural*, *extensional* or *background knowledge*) of alignment techniques [6] (e.g., alternative string matching techniques based on a confidence score [11]). Custom implementations of alignment techniques is done using Python.

Thirdly, the extent to which additional one-to-one mappings can be produced by applying disambiguation techniques on one-to-many mappings is explored. Disambiguation techniques are based on general heuristics applied on the metadata of the alignment results, similar to how Tordai *et al.* compared structural similarities between alignments [13].

Given the size of both vocabularies, evaluation of alignment results is done by means of a Gold Standard. A subset of high quality alignments between GTAA and GeoNames is produced which is used to verify the recall and precision of other alignment results. The performance of disambiguation techniques is evaluated based on a combination of precision and recall.

[8] `https://www.beeldengeluid.nl/en`
[9] `http://www.geonames.org/`
[10] `http://cultuurlink.beeldengeluid.nl/app/`

5 Alignment Results

This section describes the results of the method described in Section 4. Firstly, the differences and similarities between the used vocabularies are described as a result of an initial assessment. Secondly, the quantity and quality of the alignments resulting from the performed alignment strategy are outlined.

5.1 Initial Vocabulary Assessment

Vocabulary characteristics As stated in Section 4, the source vocabulary used in this research consists of the geographical concepts in the GTAA and the target vocabulary is the entirety of GeoNames. The central result of the performed initial vocabulary assessment is that both vocabularies are of high quality, meaning that making high quality alignments between the two is feasible. Even though both vocabularies contain geographical concepts, they nevertheless vary in a number of other characteristics. These characteristics can impact available alignment strategies for this research. Table 1 gives an overview of the different characteristics of the vocabularies. The remainder of this section describes how these characteristics impact the further stages of this research.

Table 1: Comparison of characteristics of the vocabularies

Characteristic	**GTAA Geo**	**GeoNames**
Amount of concepts	14.243	12 million
Hierarchically connected concepts	0.02%	99.88%
AltLabels per concept	0.02	1.43
SKOS quality issues	None	NA

At the time when this research was performed, GTAA had 14.243 unique geographical concepts that were approved by NISV. All concepts have had a unique concept ID assigned to them in an *rdf:about* attribute, alongside a unique *skos:prefLabel*. Of these concepts, only 287 have one or more *skos:altLabel* and only 286 concepts are hierarchically connected via a *skos:related* relation. As such, there is little to none additional information available on the concepts aside from their *skos:prefLabel*.

The entirety of GeoNames however, consists of close to 12 million unique geographical concepts. Due to limitations described in Section 8.3, only the labels present in the GeoNames daily data dump "allCountries.zip"[11] are used in this research. Similarly to the GTAA concepts, all of the GeoNames concepts have a unique concept ID and a preferred label. But in contrast to the GTAA concepts, most GeoNames concepts are enriched via the addition of hierarchical relations, alternative labels, and other data fields (e.g., population, feature class and coordinates). Practically all concepts (*99.88%*) are enriched with hierarchical relations. Of the 12 million GeoNames concepts, 6.1 million (*51%*) have one

[11] `https://download.geonames.org/export/dump/`

or more alternative labels. With a total of 17.1 million alternative labels, the concepts have an average of roughly 1.5 alternative labels per concept.

Impact on available alignment strategies Of the described vocabulary characteristics of the GTAA and GeoNames, two have major implications as to the available alignment strategies. The first being the absence of heterogeneous information about the GTAA concepts, and the second being the large amount of concepts described in GeoNames and the GTAA.

Two categories of techniques used by alignment tools are lexical and structural alignment [6]. The assessment described above concluded that the GTAA had minimal hierarchical relations between the concepts, meaning that structural alignment techniques are not be feasible. Additionally, the lack of alternative labels or other lexical labels, aside from a preferred label, used to describe the GTAA concepts limits the availability of lexical alignment techniques.

Vocabularies from the cultural heritage domain are large and sparse in nature compared to other domains, meaning that they commonly contain between 10.000 and a 100.000 concepts that are loosely connected. However, many alignment tools are attuned to produce alignments between a much smaller but tightly connected set of concepts. As such, producing alignments in vocabularies from the cultural heritage domain with these tools often times either takes too much time or simply fails [15]. Given this, the 14k geographical concepts described in the GTAA poses a problem for many of the standard alignment tools. Moreover, the 12 million concepts from GeoNames entails that an initial set of alignments can only be produced by the most basic of alignment techniques (primarily string matching). While these limitations strain the complexity of available alignment techniques, performing disambiguation techniques on the resulting set of alignments is still feasible.

5.2 Aligning GTAA to GeoNames

Baseline alignment Section 2.2 outlines the three main steps of the process of producing alignments between vocabularies. By means of exact string matching between the concepts of the two vocabularies, a baseline set of alignments is produced. The process of this exact string matching is visualised in Figure 1. For each concept in the GTAA vocabulary, a comparison is made between its *skos:prefLabel* and every *Name* in GeoNames. If the two labels are a match, the GeoNames concept ID is appended to the result set of the GTAA concept ID.

The exact string matching process splits the GTAA concepts into three different categories, which are displayed in Table 2. If only a single GeoNames concepts maps to a GTAA concept, this mapping is part of the one-to-one (OTO) mapping set. This set of 3.115 (*22%*) OTO mappings is accepted as an initial set of correct alignments [13]. The one-to-many (OTM) mapping set consists of 6.912 (*49%*) GTAA concepts, with a total of 47.905 mappings between the two vocabularies. With an average of 7 mappings per GTAA concept in the OTM mapping set, disambiguation is needed to determine which single mapping is to be used to produce a correct alignment.

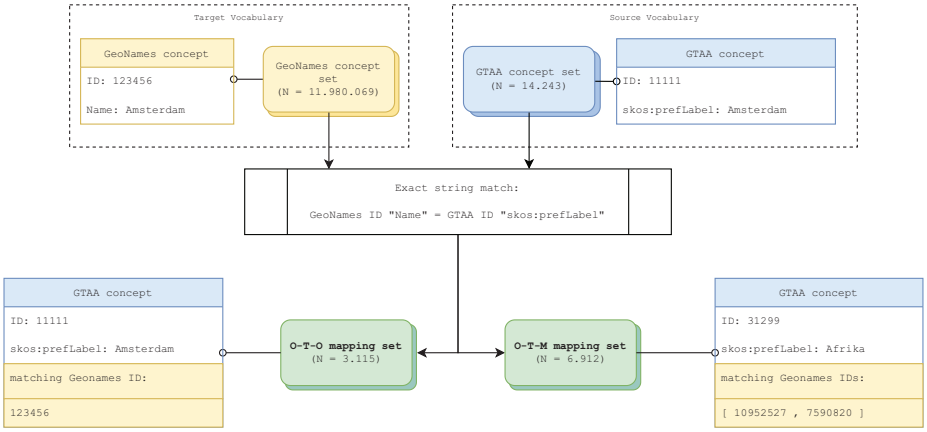

Fig. 1: Exact string matching alignment strategy outline

Table 2: Exact string matching results

Result set	**Amount of GTAA concepts**
one-to-one mapping	3.115
one-to-many mapping	6.912
no mapping	4.216

Ambiguous alignment techniques After producing a baseline set of alignments, the method proposed by Tordai *et al.* suggest that a set of overlapping alignments is produced by using more ambiguous lexical and structured techniques [13]. However, the initial assessment of the vocabularies in Section 5.1 concluded that the size and available lexical information of the vocabularies hinders the execution of these more ambiguous techniques. Given this, the overlapping results that are to be produced in this step of the process consist solely of the OTM mapping set produced via exact string matching.

Normally, exact string matching does not produce a sufficient amount of overlapping alignments on which to compare disambiguation techniques. However, the vast size of GeoNames enables the production of close to 50k overlapping alignments spread over 7k concepts. This sizeable result set in the baseline alignment step justifies the deviation from the suggested alignment procedure.

Disambiguation techniques The final step in the suggested alignment procedure is to apply disambiguation techniques on the overlapping alignments. In an effort to answer the first research question presented in Section 3, disambiguation techniques based on the geographical properties of the concepts that are part of the OTM mapping set are developed.

Ideally, these disambiguation techniques were to include comparisons between a multitude of geographical properties such as coordinates or hierarchical similarities. However, since this data was not available for the GTAA concepts, the developed disambiguation techniques were limited to the GTAA *prefLabels*

and *scopeNotes*. ScopeNotes for GTAA geographical concepts generally contain a single word description of where a concept is located. For example, 1308 GTAA concepts have the literal *"Nederland"* as their scopeNote. These scopeNotes are used in the disambiguation techniques by comparing them to the various hierarchical data available for the GeoNames concepts.

In Figure 2, the overall process used to disambiguate OTM mappings is presented. A combination of the lexical functions *Longest common substring* and *Levenshtein distance* is used to produce an initial scoring between the GTAA scopeNotes and the various GeoNames mappings. After this initial score is calculated for each mapping, the mappings with a score higher than the cut-off are passed through a filter that selects the single mapping based on a predefined criteria.

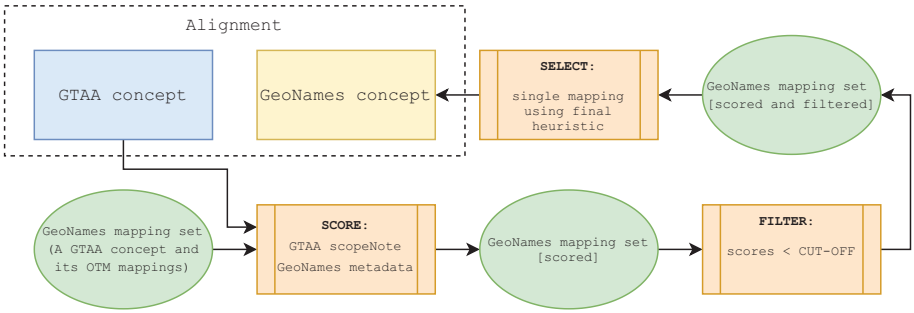

Fig. 2: Outline of the developed disambiguation technique

Table 3 describes the different variants of the developed disambiguation techniques. The scoring, cut-off and filter columns refer to the configurations of the described disambiguation technique. The scoring step has a rule-based variant and several score-based variants. The rule-based variant relies on exact matching, while the score-based variants consist of multiple weight configurations between the *Longest common substring* and *Levenshtein distance*. The cut-off for the score-based variants is manually chosen based on an initial evaluation of the scoring variant. For the filter step, *"FC.P/A"* refers to whether a GeoNames concept is considered to be a physical place (P) or an administrative division (A) and *pop* refers to filtering the mapping with the highest population. In general for each variant of the disambiguation technique, the GeoNames concept of which the hierarchical metadata resembles the GTAA scopeNote and with the highest population is selected to be the correct mapping.

Table 3: Description and performance of the various disambiguation techniques

ID	scoring	cut-off	filter	alignments	predicted correct	recall	precision	F1
1	rule	1.00	FC.P + pop	4254	3808	0.62	0.90	0.73
2		1.00	FC.A + pop	3170	2963	0.46	0.93	0.62
3		1.00	pop	4373	3891	0.63	0.89	0.74
4	score 1	0.05	FC.P + pop	5587	4470	0.81	0.80	0.80
5		0.05	pop	5865	4567	0.85	0.78	**0.81**
6		0.30	pop	1734	1441	0.25	0.83	0.39
7	score 2	0.05	FC.P + pop	4673	3675	0.68	0.79	0.73
8		0.30	FC.P + pop	4673	3675	0.68	0.79	0.73
9		0.30	pop	4894	3754	0.71	0.77	0.74
10	score 3	0.30	FC.P + pop	4839	3808	0.70	0.79	0.74
11		0.30	pop	5103	3917	0.74	0.77	0.75
12	score 4	0.30	FC.P + pop	4053	3414	0.59	0.84	0.69
13		0.30	pop	4186	3468	0.61	0.83	0.70

Evaluation of disambiguation techniques In order to evaluate the alignments produced by the disambiguation techniques, a Gold Standard of mappings between GTAA and GeoNames is developed. A small subset (N=2050) of GeoNames concepts located in the Netherlands is gathered by querying WikiData[12] for all Wikidata concepts associated with a GeoNames ID and a *BAG* ID. BAG[13] ("Basisregistratie Adressen en Gebouwen") is a dutch standard for registering addresses and buildings. The peer reviewed mappings between Wikidata, GeoNames and BAG allows us to assume these concepts to be of high quality. Of the 2050 concepts in the wikidata query result set, 1325 (*64.6%*) were linked to GTAA concepts by means of an exact string match on the prefLabels using an alignment strategy formulated in Cultuurlink[14].

Table 3 shows the performance of each of the 13 disambiguation techniques. While the lack of metadata about the GTAA concepts resulted in disambiguation techniques that were rather non-diverse in nature, consistent performance of the various configurations shows the added benefit to the production of alignments. The disambiguation techniques on average produce alignments for 4416 out of the 6912 concepts in the OTM mapping set, giving them an average recall of *63.9%*. The precision of these alignments is estimated by comparing the GeoNames/GTAA alignment to the alignment present in the Gold Standard. This precision allows for a rough prediction of how many alignments produced by the disambiguation technique are correct alignments.

In general, the stricter variants of the disambiguation technique (IDs 1,2,3 and 6) have a lower recall and a slightly higher precision compared to the more forgiving variants. This trade-off between precision and recall is a commonly observed phenomenon.

[12] `https://www.wikidata.org/wiki/Wikidata:Main_Page`
[13] `https://bag.basisregistraties.overheid.nl/`
[14] `http://cultuurlink.beeldengeluid.nl/`

Calculating the F_1 score [9], this being the harmonic mean of precision and recall, is a way to score the overall performance of precision and recall trade-offs. While it is not by any significant margin, disambiguation technique with ID "5" can be assumed to be the best performing variant given the F_1 scores.

Putting it all together The initial assessment concluded that making alignments between GeoNames and the geographical concepts of the GTAA was feasible, but with certain limitations to the available alignment techniques. The baseline set of alignments is the one-to-one mapping set of 3.115 alignments produced by exact string matching. An additional 6.912 GTAA concepts had a one-to-many mapping relation to GeoNames. Despite the limitations on available alignment techniques, the best performing disambiguation technique was able to successfully produce 4.567 additional alignments with a recall of *85%* and a precision of *78%*. Extending the baseline alignments by including the alignments produced by this disambiguation technique results in a total of 7.682 alignments between GTAA and GeoNames, which is a *147%* increase to the baseline.

6 Generous Interface Method

Generous interface design. After evaluating the alignments produced using the described method, the aligned concepts in the source vocabulary can be enriched using metadata from the target vocabulary. In other words, data that is part of GeoNames concepts can now also be displayed alongside the GTAA concepts they have been aligned with. To exemplify, the aligned GTAA concepts can now also be filtered on their population size, higher administrative division (e.g., province or state), or coordinate location. Concepts enriched with this additional metadata can then be displayed to users in a generous interface (described in Section 2.4). For this research, a prototype of a generous interfaces is developed in which videos from the Openbeelden collection[15] are displayed using geographical enriched metadata.

The goal of this interface is to investigate how such an interface addresses challenges regarding information seeking needs for experts in the cultural heritage domain. The effectiveness of this generous interfaces in addressing challenges in information seeking needs in the cultural heritage domain is evaluated using in-depth interviews with domain experts. Interviews provide limited opportunities for quantifying the effect of generous interfaces. However, this qualitative evaluation method is chosen due to the exploratory nature of the use-case, as well as time constraints for the researcher.

The interviews conducted for this research are loosely structured, with some questions prepared to allow for concrete answers to the research questions. During the interview of about an hour, the cultural heritage experts are invited to interact with the developed generous interface and discuss their experience. For example, if needed some basic search operations (e.g., "How many videos would you say there are in Limburg?") are prepared to get the experts acquainted with the interface. However, the general experience is that the experts already explored their own area of interest on the interface prior to the interview.

[15] `https://openbeelden.nl/media`

7 Generous Interface Results

A practical result of the alignment process described in Section 5.2 is that over half (*54%*) of the 14k GTAA geographical concepts are enriched with metadata from GeoNames. Available metadata includes coordinates of concepts and geographical hierarchies such as countries or provinces. Displaying cultural heritage objects that are annotated using these enriched GTAA geographical concepts could address some issues regarding information seeking needs in the cultural heritage domain. This section outlines the development of a generous interface using the alignments produced in Section 5.2. Furthermore, an initial assessment of issues regarding information seeking needs in the cultural heritage domain that could potentially be addressed using generous interfaces is given. Finally, results from interviews with cultural heritage experts about the ability of the developed generous interface to address these information seeking needs is outlined.

GeoDisplay The generous interface developed for this research goes by the title of *"GeoDisplay"*, the landing page for this interface is displayed in Figure 3. A screencast for the interface is available[16], the interface itself is also accessible online[17]. The cultural heritage objects that are displayed in the interface are public domain videos from Openbeelden published by the NISV. The coverage of locations in these videos has been annotated using GTAA geographical concepts. For this prototype, the videos have been restricted to those annotated using GTAA concepts located in the Netherlands. This restriction to filter for concepts located in the Netherlands has been made possible by the additional metadata generated from the alignment process. Another example of how the interface uses the enriched metadata is by displaying the places that Openbeelden videos are annotated with on an interactive map, which requires exact coordinates. Additionally, the GTAA geographical concepts can be filtered hierarchically based on which province they are located in.

On Figure 4, an annotated screenshot from an Openbeelden video displayed on GeoDisplay is visible. Users are able to navigate between the videos primarily through geographical relations. Users can choose to use keyword search to find a specific and known place, or explore the Openbeelden video collection through a map with markers as seen in Figure 5. Once a video is selected, other videos related to the selected video by year, topic or place can be explored.

In combination with the ability to explore the context of the collection items through the functionalities described above, the home page of the interface (Figure 3) provides a high level overview of the items in the collection across various dimensions like space, time and common topics. Because the interface adheres to these principles of voluntarily providing additional information (overviews, samples, context) it can be regarded as a generous interface [16].

Information seeking issues The issues regarding information seeking needs in the cultural heritage domain identified by [2] assume information tasks to be classified in six different categories. These information task categories are adapted from [7]. Three of these task categories are relevant for classifying search behaviour of cultural heritage experts: Fact Finding, Information Gathering, and Keeping Up-to-date. Of these three categories, the majority of tasks done by cultural heritage experts fall under Information Gathering, followed by Fact Finding. Keeping Up-to-date is not observed to be a common task for cultural heritage experts.

[16] `https://youtu.be/hUM-jL8Q83Y`
[17] `https://geointerface.herokuapp.com/`

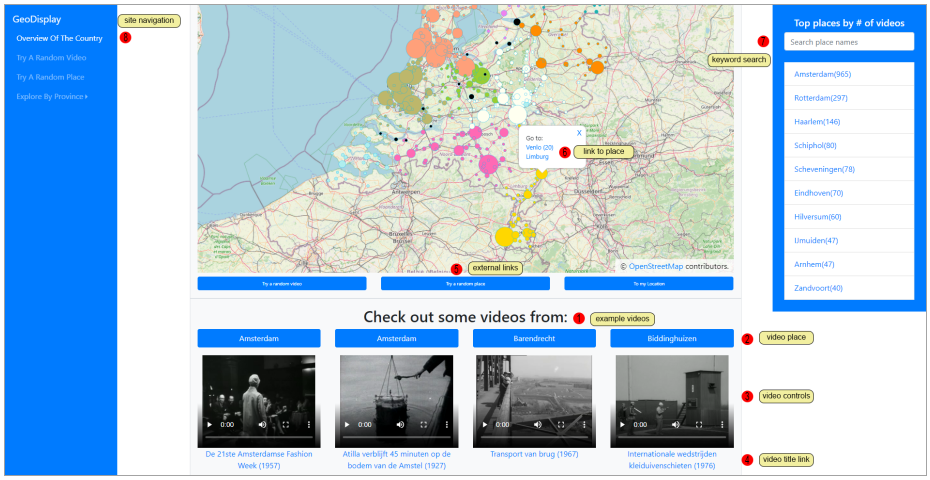

Fig. 3: Annotated home page from GeoDisplay

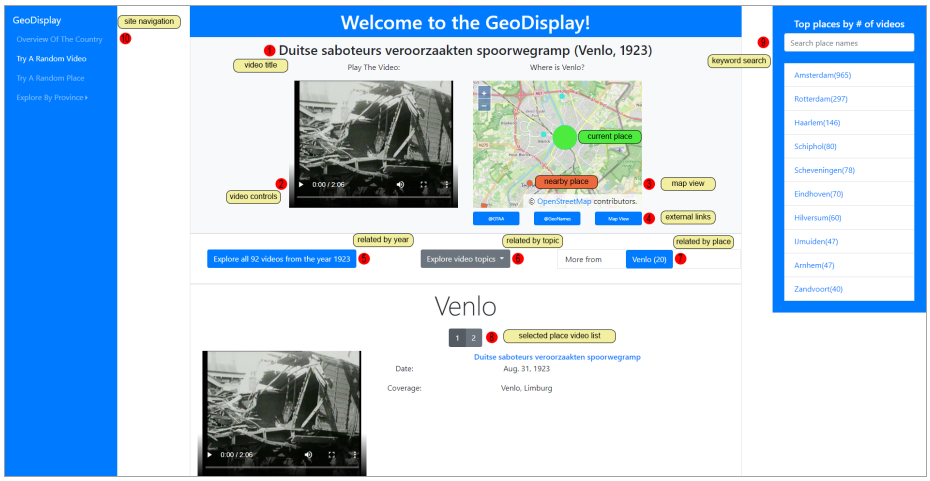

Fig. 4: Annotated video page from GeoDisplay

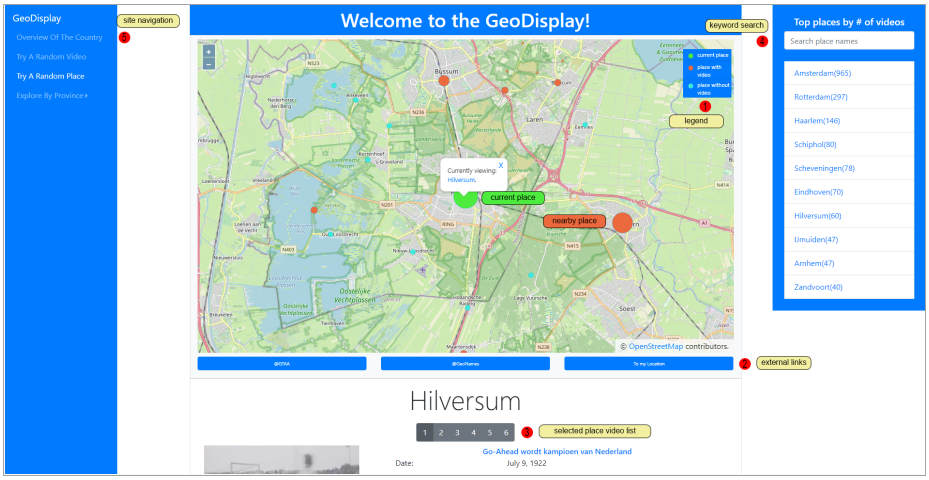

Fig. 5: Annotated place page from GeoDisplay

Fact Finding occurs when a specific piece of information is needed, causing the cultural heritage expert to do goal oriented search [2]. Common issues regarding Fact Finding in the cultural heritage domain occur in query formulation. Cultural heritage experts experience difficulty in formulating search queries in such a way that an appropriate amount of results is returned. If the result set is too large, the cultural heritage expert would have to spend extensive manual effort to filter through the results. If the result set is too small, the factual goal of the cultural heritage expert may not be among the results. This issue of formulating queries of the right level of complexity is likely to occur when the expert formulating the query is not familiar with the underlying vocabulary [2]. Since a goal of generous interfaces is to provide an entry point into a collection [16], the interface should also be intuitive enough to be operated by people unfamiliar with the collection. As such, an opportunity generous interfaces offer is to provide enough support to cultural heritage experts in the formulation of queries into the collection.

Information Gathering is when multiple search tasks are performed in order to carry out an overarching goal (e.g., writing a blog post) [2]. Given its broad spectrum, Information Gathering is further defined into a number of sub-tasks. These sub-tasks can range from comparing information between different collections to exploring different aspects of any given topic. Despite this broad spectrum of tasks, a unifying issue for cultural heritage experts performing these tasks is that many of them are not supported by current tooling. It stands to reason that generous interfaces should be able to support a larger amount of high level search activities done by cultural heritage experts compared to standard search interfaces.

Keeping Up-to-date contains the search tasks that are performed by cultural heritage experts to explore novel or otherwise relevant information within a collection [2]. The contrast with the previous two task categories is that Keeping Up-to-date tasks are not necessarily goal driven, meaning that the cultural heritage expert is not looking for specific information. Keeping Up-to-date is done by either actively searching for novel information in a collection, or by passively receiving updates. One of the reasons that

Keeping Up-to-date is not observed to be a common task for cultural heritage experts is because only a low number of tools support similar features. Providing users with an overview of "What's new" or "Have you seen this" is in line with the principle of generosity of generous interfaces [16]. As such, generous interfaces could facilitate an increase in tools that support Keeping Up-to-date features.

Interface evaluation As described in Section 6, interviews with cultural heritage experts were performed in order to evaluate the extent to which the developed generous interface is addressing issues regarding information seeking needs in the cultural heritage domain. A total of three interviews were performed with an average length of an hour. In these interviews, the cultural heritage experts elaborate on how they would be using the developed interface for their research or other cultural heritage related activities. The interviews also provided further verification of the existence of the information seeking issues identified by [2]. In line with the categorisation of information tasks in which the information seeking needs are defined, the evaluation of the developed interface can also be viewed across these categories.

Fact Finding: GeoDisplay allows cultural heritage experts to explore the Openbeelden collection based on the geographical coverage of the collection objects. The cultural heritage experts had not previously explored the Openbeelden collection, meaning that exploring the collection via queries would not be ideal [2]. When presented with the interface, the cultural heritage experts generally expressed interest in videos from a specific city or region. Because the experts were familiar with the region in which they were interested, locating it on the map present in the interface was possible. The cultural heritage experts evaluated the navigation of this map to a specific location they were interested in as easy and intuitive. Navigating this map can be viewed as an alternative way to build a query to the collection, given that a cultural heritage expert wants to collect a specific subset of the collection. One observed aspect of exploring the interface in this way is that the map not only provides information on what is in the collection, it also provides information on what is not in the collection. Without the need of multiple query inputs, cultural heritage experts could find out whether or not the collection contains videos covering a place by checking if the place has a marker placed on it.

Information Gathering: Being the most prominent search task performed by cultural heritage experts, findings towards the improvement of tool support for Information Gathering tasks have significant merit. The developed interface was observed to support some higher level information tasks that were not supported by standard search-based interfaces. The general map overview presented by the interface was observed to provide an entry point to perform exploration based information gathering tasks. After navigating the map to select a place they were interested in, cultural heritage experts were presented with the option to explore places located near their selected place. One cultural heritage expert was observed to navigate to Amsterdam, following which they switched their attention to nearby places such as "Rokin" or "Vondelpark" using the map functionality as visible on Figure 5. This meant that they were exploring the contents of the Openbeelden collection without the need of a constant formulation of relevant queries.

Selecting a place using the interface automatically generates relevant suggestions in the form of nearby places, which was made possible by the enriched geographical metadata produced by alignments. Cultural heritage experts also noted how the various related videos accessible via 5, 6, 7 and 8 in Figure 4 helped them to explore relevant information about various topics. For example, one cultural heritage expert stated *"I want to see more related videos about swimming"*, another wanted to know *"What*

other videos are there about Rotterdam before WW2". These questions are not typically supported by keyword search [2]. However, the developed generous interface was able to successfully help the cultural heritage experts find answers to these questions by providing related search results.

Keeping Up-to-date: In line with previous findings, Keeping Up-to-date was not a commonly performed task by the cultural heritage experts. One expert noted how it was very difficult to keep up with changes made to the various collection interfaces themselves. This seems reasonable, given that cultural heritage experts generally use many different sources accessible through these kinds of interfaces. The "What's new" aspect of Keeping Up-to-date could not be evaluated using this interface, given that the cultural heritage experts only interacted with the interface on a single occasion. However, the "Have you seen this" aspect was observed to be present in the interface. The cultural heritage experts would come across new videos in a serendipitous manner. Examples of this include conversations like *"Ah Leek (village in the Netherlands), I like that, that is what I am working on at the moment"* or *"I also came across something that made me really happy, there was this cinema ... There were pictures of it after it was destroyed, but this was the first time I have seen in before that"*. Enabling such spontaneous findings are a major motivator behind the principles of generous interfaces [16]. As described above, the cultural heritage experts were observed to come across videos that were interesting to them in a serendipitous manner. This provides further evidence that generous interfaces are able to fulfil *keeping up-to-date* information needs of cultural heritage experts.

The evaluation of the developed generous interface has shown that a significant number of issues regarding information seeking needs in the cultural heritage domain can be addressed by providing access to collections through such interfaces. By supporting alternative ways to formulate queries and by providing relevant related search results, the developed generous interface was able to assist cultural heritage experts with both Fact Finding and Information Gathering tasks.

8 Discussion

This section relates the findings to additional (practical) contributions of the research beyond the research questions. Additionally, some findings of the generous interface evaluation not directly related to the results presented in Section 7 are discussed. Finally, some encountered limitations to the various aspects of this research are discussed.

8.1 Practical use of produced alignments

In line with previous research on alignments [13], the baseline alignments produced by this research could be accepted into the GTAA without too many incorrect alignments being introduced. However, accepting the alignments produced by the disambiguation techniques without modification may not be in the best interest of institutes like the NISV as it would introduce a predicted average of about 800 incorrect alignments. This does not mean that the alignments produced by the disambiguation techniques have no value to institutes like the NISV, as the results have the majority of obviously incorrect ambiguous mappings filtered out. Another option the NISV could consider is to accept the alignments produced by the disambiguation techniques, with a relevant annotation describing the provenance of the alignment. This would allow users of the data to take the creation process of the alignments into consideration when drawing their conclusions.

8.2 Generous interface

The evaluation of the developed generous interface provide insight on how the geographically enriched GTAA concepts could be used to address issues regarding information seeking needs in the cultural heritage domain. Furthermore, the interviews shed some interesting lights on the relation between generous interfaces and cultural heritage experts. The cultural heritage experts agreed that interfaces such as these are mostly used as an entry point for research, and that they should be used next to other interfaces and not instead of. This is in line with the motivation behind generous interfaces [16]. The use of multiple interfaces next to each other gives cultural heritage experts the ability to cross reference various sources in order to *"fill in the blanks"* or *"correct errors in the metadata"*. This need for the manual comparison of different sources was identified as a problem in information seeking tasks [2]. Nevertheless, the cultural heritage experts were of the opinion that this was a vital part of their job and that their ability to do so set them apart from the more casual users of such interfaces. However, this of course does not take away the need for this identified problem to be addressed as this would benefit the cultural heritage experts the most.

Another identified issue for cultural heritage experts is that creating queries that return a result of the right size is difficult when they are unfamiliar with the vocabulary used to describe the collection [2]. However, the evaluation concluded that for many of the information seeking tasks that the cultural heritage experts were unable to perform the issue lay not with the interface but with the granularity and quality of metadata. Generally speaking, the GTAA geographical concepts associated with the Openbeelden videos are cities or villages. However, the cultural heritage experts expressed interest in videos based on places such as buildings, streets or landmarks. The cultural heritage experts proposed that adding this more precise geographical data could be achieved by allowing parties outside of the NISV to contribute this metadata. One cultural heritage expert suggested this could be done by asking local film enthusiasts for their input, another proposed that the developed interface could be used by history students to research and add this information.

The developed generous interface only addressed a single use-case in the form of displaying the Openbeelden videos using enriched geographical data. However, during the evaluation it became apparent that there is further demand for interfaces such as these among cultural heritage experts. When asked what other geographical metadata displayed in a generous interface would be able to assist a cultural heritage expert, the idea arose to overlay the map with local tax data to gauge the income levels of different neighbourhoods over time. Another idea that came to the mind of a cultural heritage expert was to highlight the Openbeelden videos along popular historical city centres. One can imagine that practically every cultural heritage expert would have different needs when it comes to connecting data sources. While it is impossible to create a generous interface for each imaginable information need, having linked data of sufficient quality so that it can easily be aligned with other sources would enable many opportunities.

8.3 Limitations and future work

Some aspects of the performed research method had certain limitations attached to them. This section outlines the most important limitations that resulted from this.

Alignment limitations As stated in Section 5.1, the size of GeoNames proved to be too large to handle a number of alignment strategies. However, it also had some

other unforeseen practical implications. A standard format to use when dealing with linked data is some representation of the Resource Description Framework (RDF). While the individual GeoNames records can easily be obtained as RDF, obtaining the entire dataset in RDF is only possible as a single 18GB export. This research lacked the resources to perform the proposed methodology on such a file. As a compromise, the .CSV export "allCountries.zip" (1.5 GB) of GeoNames was used. However, this lacks some critical information compared to the .RDF such as labels in alternative languages.

Interface limitations During the evaluation of the generous interface, two aspects of the interface were found to be lacking in development. Firstly, poor optimisation of the interface introduced long loading times on slow devices. This is because the interface sends all data that comes as a response to a request in a single message. While this did not impact the available functionality of the interface, the increased loading times may have hampered the motivation of cultural heritage experts to explore the collection. Secondly, the fact that the interface was largely non-responsive caused the display to be bugged on smaller screens. The bugging caused certain buttons or text to be hidden over each other. Luckily, this could manually be fixed during interviews by zooming out in the browser. However, this could hamper outside parties in their verification of the findings regarding the developed prototype.

Future work could expand upon this research by adopting this methodology using smaller but more detailed geographical vocabularies. This would enable a more in-depth research into which aspects of geographical disambiguation techniques are effective in producing alignments. Additionally, geographically enriched metadata could be used in the production of generous interfaces that address other use-cases in the cultural heritage domain. This would allow for a broader evaluation of generous interfaces in the cultural heritage domain.

9 Conclusion

The evaluated one-to-many disambiguation techniques based on the geographical metadata of both GTAA concepts and GeoNames concepts produced a sizeable extension to the baseline set of alignments produced by exact string matching. Using this set of alignments, a generous interface displaying Openbeelden videos was developed. The interface was evaluated by experts from the cultural heritage domain. This evaluation concluded that in the use-case for which the interface was designed, the inclusion of additional geographical metadata enabled viable solutions towards a number of issues regarding information seeking needs in the cultural heritage domain.

acknowledgements This research was done as part of an internship program within the NISV, with the kind and ever-present assistance of Jesse de Vos.

References

1. Amin, A., Hildebrand, M., Van Ossenbruggen, J., Hardman, L.: Designing a thesaurus-based comparison search interface for linked cultural heritage sources. pp. 249–258 (02 2010). https://doi.org/10.1145/1719970.1720005
2. Amin, A., Van Ossenbruggen, J., Hardman, L., Nispen, A.: Understanding cultural heritage experts' information seeking needs. pp. 39–47 (01 2008). https://doi.org/10.1145/1378889.1378897

3. Boer, V., Wielemaker, J., Gent, J., Hildebrand, M., Isaac, A., Van Ossenbruggen, J., Schreiber, G.: Supporting linked data production for cultural heritage institutes: The amsterdam museum case study (05 2012). https://doi.org/10.1007/978-3-642-30284-8$_5$6
4. Broughton, V., Hansson, J., Hjørland, B., López-Huertas, M.J.: (2005), `http://hdl.handle.net/10150/105851`
5. Euzenat, J., Shvaiko, P.: Ontology matching: Second edition (10 2013). https://doi.org/10.1007/978-3-642-38721-0
6. Isaac, A., Wang, S., Zinn, C., Matthezing, H., Meij, L., Schlobach, S.: Evaluating thesaurus alignments for semantic interoperability in the library domain. Publ. in: IEEE Intelligent Systems 24 (2009), 2, pp. 76-86 **24** (03 2009). https://doi.org/10.1109/MIS.2009.26
7. Kellar, M., Watters, C., Inkpen, K.: An exploration of web-based monitoring: Implications for design. pp. 377–386 (01 2007). https://doi.org/10.1145/1240624.1240686
8. Mader, C., Haslhofer, B., Isaac, A.: Finding quality issues in skos vocabularies (06 2012). https://doi.org/10.1007/978-3-642-33290-6$_2$5
9. Rijsbergen, C., Lalmas, M.: Information calculus for information retrieval. JASIS **47**, 385–398 (05 1996). https://doi.org/10.1002/(SICI)1097-4571(199605)47:5¡385::AID-ASI6¿3.0.CO;2-S
10. Shvaiko, P., Euzenat, J.: Ontology matching: State of the art and future challenges. Knowledge and Data Engineering, IEEE Transactions on **25**, 158–176 (01 2013). https://doi.org/10.1109/TKDE.2011.253
11. Stoilos, G., Stamou, G., Kollias, S.: A string metric for ontology alignment. vol. 3729, pp. 624–637 (11 2005). https://doi.org/10.1007/1157462$0_4$5
12. Suominen, O., Mader, C.: Assessing and improving the quality of skos vocabularies. Journal on Data Semantics **3** (03 2014). https://doi.org/10.1007/s13740-013-0026-0
13. Tordai, A., Van Ossenbruggen, J., Schreiber, G.: Combining vocabulary alignment techniques. pp. 25–32 (01 2009). https://doi.org/10.1145/1597735.1597741
14. Tordai, A., Van Ossenbruggen, J., Schreiber, G., Wielinga, B.: Aligning large skos-like vocabularies: Two case studies. pp. 198–212 (05 2010). https://doi.org/10.1007/978-3-642-13486-9$_1$4
15. Van Ossenbruggen, J., Hildebrand, M., Boer, V.: Interactive vocabulary alignment. vol. 6966, pp. 296–307 (09 2011). https://doi.org/10.1007/978-3-642-24469-8$_3$1
16. Whitelaw, M.: Towards generous interfaces for archival collections. International Council on Archives Congress **2012** (01 2012). https://doi.org/10.3828/comma.2012.2.13
17. Whitelaw, M.: Generous interfaces for digital cultural collections. Digital Humanities Quarterly **9** (2015)
18. Wigham, M., Melgar, L., Ordelman, R.: Jupyter notebooks for generous archive interfaces. pp. 2766–2774 (12 2018). https://doi.org/10.1109/BigData.2018.8622203