The benefits of linking Thesauri for internal users
A Netherlands Institute for Sound & Vision case study

TIM DE BRUYN, Vrije Universiteit

ABSTRACT

In this research we look at the Netherlands institute for Sound & Vision (NISV) as a case study to find out what the advantages and possibilities of linking their thesaurus to an online knowledge base are. NISV manages an audio-visual collection concerned with the Dutch media cultural heritage. They have chosen for Wikidata as the external dataset to enrich their collection with, through links with an internal vocabulary, GTAA. This research is part of a broader research into the advantages of using linked data at NISV. This part of the research focuses on the internal users while the other part focuses on the external users. In this research we look into the existing data contained both within the thesaurus of NISV and Wikidata. A statistical analysis is conducted to answer questions on the completeness and richness of the data. Also, several internal users of different departments of NISV are interviewed in an effort to extract use cases for an enhanced collection. Using the interviews, use cases are set up and (conceptual) prototypes are built to satisfy said use cases. The (conceptual) prototypes are then evaluated for their usefulness and added functionality. A comparison between the internal and external users and their different wishes is drawn. This research shows the differences and similarities between the existing thesaurus and Wikidata. It brings forth three use cases and lays a down a framework for future work in similar case studies.

Additional Key Words and Phrases: Linked Data, Wikidata, Digital thesauri, Cultural heritage

1 INTRODUCTION

NISV is an audiovisual archive located in Hilversum. It concerns itself with the Dutch audiovisual heritage of national documentaries, movies, music, radio and television programs. In partnership with other Dutch organizations that concern themselves with said Dutch heritage they developed the Communal Thesaurus for Audiovisual Archives (GTAA). This thesaurus allows NISV to accurately characterize audio-visual material. The thesaurus is to be enriched with data from the Wikidata dataset.

Wikidata is a free, collaborative, multilingual, secondary database, collecting structured data to provide support for Wikipedia, Wikimedia Commons, the other wikis of the Wikimedia movement, and to anyone in the world[1]. This data from Wikidata is to be manually linked to the GTAA. The data from the Wikidata dataset contains more information per subject. The Wikidata dataset contains: persons, objects, geographical data, concepts, company names etc. Since the data from Wikidata is richer than the data currently contained in the GTAA, it would allow NISV to set up more extensive use cases if these datasets were to be linked. NISV's goal is to enrich their thesaurus for internal use as well as for external audiences. In order to know what data is important for their own use cases analysis needs to be carried out. The current data contained within the GTAA has to be analyzed to find out what data it is lacking. After that we can see if this data can be supplied by Wikidata. In this research, conducted with co-researcher John Brooks, we aim to bring answers to questions on the subject of alignments between the current GTAA and Wikidata. Whereas my co-researcher focuses on use cases for external audiences, this research focuses on the internal users and their use cases. With the help of interviews with the internal users concrete use cases can be set up. These use cases, in turn, lead to prototypes being built to satisfy the use cases. This research brings insights on the analyzed data and brings forth concrete use cases and prototypes to satisfy these use cases. The problem of what to do with the extra information after linking a thesaurus is not a problem exclusive to NISV. Looking outside the frame of this particular case study, the setting up of use cases also provides a framework for future researches in similar case studies. With its unique combination of a data driven and user driven

[1] https://www.wikidata.org/wiki/Wikidata:Main_Page

approach this research provides in-depth insight in how use cases for an enriched thesaurus are derived.

2 RELATED WORK

Previous researches have been conducted on the subject of Wikidata and its use for enhancing existing thesauri. Research has also been conducted on the quality of the data contained in the dataset of Wikidata. In research describing what Wikidata is and what it entails Vrandečić & Krötzsch (2014)[1] state that Wikidata's goal is to allow its data to be used both internally and in external applications, such as the external thesauri GTAA. They describe it as a community controlled database. Erxleben et al. (2014)[2] have conducted research on Wikidata and its possibilities for being connected to the linked data web. They describe how Wikidata is nowadays already linked to multiple external datasets. These existing links range from the ISSN dataset, which identifies all journals and magazines etcetera, to more highly specialized databases, such as the database of North Atlantic hurricanes. Färber et al. (2016)[3] looked into the quality of the most noteworthy large knowledge databases, which they describe as knowledge graphs. On the knowledge graph this research is concerned with, namely Wikidata, extensive research show that multiple quality aspects of the data contained within Wikidata were the highest of any of the noteworthy knowledge graphs. Examples of quality aspects relate to: Accuracy, Trustworthiness, completeness, interoperability etc. This research could show that the data contained within Wikidata is indeed the best suitable data for enriching the GTAA. Thornton et al. (2017)[4] also conducted research on Wikidata's possible role to serve as repository for international organizations concerned with digital preservation. They concluded that Wikidata had the advantage of being structured, queryable and computable. Another advantage Wikidata has is that it's multilingual. Since Wikidata also had a Dutch version it supports alignments to the GTAA better.

Debevere et al. (2011)[5] have conducted research in the linking of thesauri via linked data as to improve the metadata of thesauri. This improved metadata then further allows better retrieval of information on search queries. They developed an algorithm that automatically linked data categories of DBpedia to another media thesaurus keyword thesaurus used to annotate archived media items at the Flemish public service broadcaster (VRT). It returned acceptable results as 91.42% of the category "Person" and 89.94% of the category location were correctly linked. Tordai et al. (2007)[6] tried to provide a systematic approach to build a large semantic culture web. They did so by making clear to heritage institutions what they need to do to make their collections fit for becoming a part of this semantic culture web. In short they advise to use to use the paradigm of open access, open data and open standards. They conclude that collection owners should be provided with the necessary support facilities. In a similar case study to this one Kobilarov et al. (2009)[7] looked into the use of linked data in a case study at the BBC. They linked two separate platform of the BBC to each other via DBpedia. They concluded that in the end the extra links and data available to the end users benefitted them. They did however agree that much more progress in the linking engine and DBpedia in particular were needed to achieve full optimization. Caracciolo et al. (2011)[8] performed a case study for the use of linked data at the AGROVOC multilingual thesaurus maintained by the Food and Agriculture Organization of the United Nations. They had a much more explorative research as in the end they only advised on the changes needing to be made to the current thesaurus for becoming accessible to a linked data web. De Len et al. (2011)[9] performed a similar use case study. They however focused on Spanish geographical data and how to rewrite it to make it benefit from the linked open data principle. Lastly, Neubert (2009)[10] also performed a similar case study to the case study performed in this research. He describes the linking of a thesaurus for economics to DBpedia and other large thesauri using SKOS/RDF methods. In the end the linking to DBpedia returned a high number of unsuccessful matches. These unsuccessful matches are contributed to simple derivations in the compared strings and the fact that a significant number of economic concepts did not exist in DBpedia yet. They also describe a method for checking for inconsistencies between linked datasets using SPARQL. This is a method that could also be useful for this research.

What we learn from all this related work is that most researches work with a data driven approach. We are also using this approach in this research, but are combining it with a user driven approach as we think this could better lead to relevant use cases. Most researches conclude with a statement on the usefulness of linked data and Wikidata in particular. This a conclusion this research agrees with and tries to back up with statistics and use cases.

3 RESEARCH QUESTIONS

Based on the motivation in the introduction and the related work, in this research we investigate the following research questions.

- What use cases can be derived from an enriched GTAA?

To best answer this question research is conducted into the data currently contained within the GTAA. This research provides a clear overview of what data has to be enriched to set up the use cases. After the use cases are set up it can be determined what alignments in the datasets have to be made to satisfy them.

- How do the wishes of internal and the external audiences differ?

As stated before this research focuses on the wishes of the internal users at NISV, whereas John Brooks focuses on the wishes of the external audience. Both our last research questions are a combined effort. We used each other's data to be able to answer the respective research questions.

3.1 Methods

This research answers its research questions using the following methods.

3.1.1 Analysis of the data within internal sources, external sources and on the richness of the theoretical link.

The data currently contained within the GTAA is analyzed to see what data it exactly contains. This also allows us to see what data it lacks. The analysis of the data is performed using existing tool currently available at NISV and via Wikidata. The data currently contained within the external sources is also analyzed. Together with the findings on the lack of certain data currently contained within the GTAA, it is checked how suitable the data from the external sources is to link to the GTAA. The data that was lacking in the current GTAA is theoretically added from external sources. This allows us to create statistics on the richness of the theoretical new dataset. These analyses allow us to create a clear overview on the data of both the datasets.

3.1.2 Interviews with the internal users and setting up use cases.

After the analysis of the data of both internal and external sources interviews with the internal users are conducted. These interviews are conducted with at least a total of 5 people from relevant departments. The interviews are divided into two parts. The first part of these interviews contains the showing of all gathered statistics on the datasets. The second part contains open questions about possible use cases. The goal of having open questions at the end of the interview is the sparking of new ideas by first walking through the possibilities of the data in the external sources. Together with the data from the analysis and the data gathered in the interviews use cases can be set up. With the use cases, prototypes can be set up which in turn satisfy the users' needs. This allows us to answer the first research question.

3.1.3 Patterns in the use cases.

After research has been carried on the internal audiences' wishes, analysis on the patterns found in the data satisfying these use cases is conducted. The goal of this analysis is to create a framework of conceptual queries that could be used in future alignments between other similar datasets. It also aids future research on the subject.

3.1.4 Exploration and working out use cases in form that best suits them.

Depending on the use cases retrieved earlier the best way to enable said use case is worked out. We determine if it is possible to translate the use cases into a working prototype or if it is best to be thought of in a conceptual manner.

3.1.5 Evaluation of the use cases.

After the use cases are set up two evaluation rounds are taking place. The goal of the first evaluation round is to determine if the worked out use cases are what the internal users envisioned. After the first evaluation round it is also possible for the users to express their desired changes to the worked out use cases. The second evaluation round serves as a medium to express the perceived usefulness of the finalized use cases.

3.1.6 *Comparison of differences between the wishes of internal and external audiences.*

In order to answer the last research question, the findings of this research is compared to that of my co-researcher John Brooks (2018). As stated before, my co-researcher focused on the external audiences while this research is concerned with the internal users. Analysis on both parts of the spectrum allows us to draw conclusions on the differences in the wishes between the two audiences.

4 ANALYSIS OF THE DATA

4.1 Analysis setup

For the analysis of the data currently contained within both the GTAA and Wikidata itself we used the Wikidata SPARQL endpoint tool[2]. We set up the statistical analysis with three goals in mind. First, we wanted to create an overview of the data currently contained within the GTAA. Second, we wanted a more global overview of the data contained within Wikidata. Lastly, we draw a comparison between the different datasets to see if the data contained within the GTAA is in line with the global data.

The goal of this analysis is to see if the data currently already linked is useful for NISV. The total amount of persons in the GTAA that is to be linked with Wikidata is 123,152. For the generated statistics on the data contained within the GTAA we looked at a subset of 10,350 persons. This subset was chosen on the premise that this was the subset of persons that were already linked. Therefore, our subset of already linked data is currently a sample size of roughly 8.4% of the total. If we look at the minimum required sample size at a confidence level of 95% and a margin of error of 1% we see that our subset should at least contain 8,909 persons. The subset meets this requirement.

The subset of persons and their properties were retrieved. Henceforth, this subset is called subset A. Of the persons retrieved not all properties were chosen for analysis. A selection of properties was made based on relevancy for NISV. This was selection was made in consideration with internal users at NISV. After talking with said users about all the existing properties they thought these to be most relevant. The conclusions on the statistical analysis are strictly bound to these properties as other properties could return different statistics. The chosen properties are as follows:

- Name
- Gender
- Awards received
- Educated at
- Member of political party
- Nominated for
- Place of death
- plays sport

To retrieve subset A and their respective selected relevant properties we used SPARQL queries to restrict all the persons contained within Wikidata to only those who already have a GTAA identifier as a property. Of those persons we retrieved their Wikidata ID, GTAA ID, name and the relevant properties in table format. This allowed us to look further into the subset.

For the analysis of the data contained within Wikidata as a whole we looked at a subset of 100,000 persons. This subset of persons was picked by random selection by the SPARQL endpoint tool. For this subset we retrieved the same properties. For this subset we once again look at the minimal required sample size at a confidence level of 95% and a margin of error of 1%. The total population at the most recent Wikidata data dump[3] is 4,310,706. This means that our subset should contain at least 9,583 persons. The subset meets this requirement. The subset is named subset B.

We expected the outcome of our statistical analysis to show some sort of relevant difference. This relevant difference could show itself in differences in nationality, place of birth and place of death. A difference in these properties was to be expected as the GTAA is mainly concerned with Dutch persons and Wikidata as a whole is an internationally diverse set of persons.

[2] https://query.wikidata.org/

[3] https://denelezh.dicare.org/gender-gap.php

4.2 Statistics

For the statistics we focused on the twenty returned properties with the highest occurrence rate. The full statistics for subset A and subset B can be found in an online file sharing platform[4].

When first looking at the gender difference we cannot compare subset A to subset B. This is attributed to the fact that the SPARQL endpoint tool had its limitations. It could in fact not generate statistics on the gender distribution in subset B. This is because the occurrence rate of the property "Gender" is too high. For this comparison we once again look at the most recent data dump. This gives us the following gender distribution for subset A in table 1 and the whole population in table 2.

Table 1. Gender distribution for subset A

Gender	Count	Percentage	Rank
Male	3,060,702	81.77%	1
Female	681,697	18.21%	2
Other	517	0,01%	3
Total	3,742,916	100.00%	

Table 2. Gender distribution of the whole population

Gender	Count	Percentage	Rank
Male	3,060,702	81.77%	1
Female	681,697	18.21%	2
Other	517	0,01%	3
Total	3,742,916	100.00%	

Here we clearly see that subset A is very much in line with the whole population. Only a small difference of 3.3% in male properties is denoted.

When comparing the statistics of only subset A and B several things immediately come forward. We see that subset A contains more entries referring to Dutch properties. This can best be seen in the statistics on "Awards received", "Educated at", "Member of political party" and "Place of death". We also see that subset A contains more media oriented properties. This can best be seen in the statistics on "Awards received". A noteworthy sighting is that both subset A and subset B contain mainly media oriented properties for the statistics on "Nominated for".

In conclusion, we expected the data of the two datasets to differ in certain aspects. Our analysis confirmed this expectation. The GTAA is focused on Dutch persons who have a connection to the fields of television, movies, radio, theater and music. The global data features more fields than that. We did however learn that the gender distribution is almost similar.

4.3 Richness of the data

In our statistical analysis we also looked into the richness of the data contained in Wikidata. More specific, for subset A we looked at the properties in figure 1 and some other properties that proved more relevant for NISV. Using these properties, we produced the statistics as seen in table 3 on the next page.

In this table we see the occurrence rate both in exact values and percentages of the total. In the same table we also see the count of multiple entries. These multiple entries denote that a person has multiple entries for one property. For occupation or country of citizenship this is straightforward as a person can easily have multiple, but for properties like date of birth and place of death it is less so. The explanation for multiple entries in these properties is that when it is uncertain what the correct only property is, multiple are given. Both these entries could be correct but the only true entry remains uncertain.

These statistics on the richness of the data create a clear image for the usefulness of the data. For example: Wanting to use the data to analyze different pseudonyms will probably not be very useful as in this subset only 1.2% of the persons have a Pseudonym property.

All the produced statistics on both the differences between the two subsets and the richness of the data contained within subset A are used in the interviews.

If we compare this to a recent research by Klein et al. (2016)[11] we can determine the meaning of our results. As part of their research they looked into property coverage of Wikidata over the years. Their results can be seen in table 4 on the next page.

We see that the biggest differences lie in the properties: "Date of birth", "Citizenship", "Place of birth" and "Occupation". The other properties are almost similar

[4] https://figshare.com/articles/Statistics/6859595

Table 3. Occurrences of properties aligned persons GTAA & Wikidata

	Occurrences (OC)	**% of total**	**Multiple entries (ME)**	**ME % of OC**	**ME % of total**
Name	10,350	100.00%	0	0.00%	0.00%
Gender	10,294	99.46%	0	0.00%	0.00%
Birth name	839	8.11%	12	1.43%	0.12%
Pseudonym	124	1.20%	17	13.71%	0.16%
Date of birth	10,040	97.00%	63	0.63%	0.61%
Place of birth	7390	71.40%	41	0.55%	0.40%
Date of death	3644	35.21%	38	1.04%	0.37%
Place of death	2776	26.82%	17	0.61%	0.16%
Occupation	9988	96.50%	5581	55.88%	53.92%
Country of citizenship	9976	96.39%	493	4.94%	4.76%

Table 4. Property coverage of Wikidata

	17-09-2014	**03-01-2016**
Gender	95.3%	96.5%
Date of birth	57.6%	71.7%
Date of death	28.6%	36.1%
Citizenship	42.8%	58.2%
Place of birth	24.0%	30.5%
Ethnic group	0.3%	0.6%
Field of work	n/a	0.3%
Occupation	n/a	58.7%
At least one site link	99.6%	98.1%

in richness. What we can conclude from this is that subset A has a higher coverage of properties in terms of occurrence rate than the whole of Wikidata in 2016. This in turn tells us that the richness of our sample size is rather high.

INTERVIEWS

A series of interviews with internal users of different departments of NISV is conducted. In total, five internal users are interviewed. The interviews are either conducted in a one-on-one setting or with two users from the same department simultaneously. The interviews each last an hour. The goal of these interviews is to result in use cases for the enriched dataset. The interviewee is shown the extra data contained within Wikidata. The interviewee is also shown the results of the earlier analysis. After, the interviewee is asked if he/she based on this data sees any concrete use cases for when more data gets linked. The departments of the different users interviewed are as follows: Intake, information management and Knowledge & Innovation.

In all the interviews the wish for a copyright expiration alert on the work of persons in the GTAA came forward. Even without looking at the produced statistics this proved a long desired wish. For the departments of intake and information management this proved also to be the only use case. For the department of Knowledge & Innovation two other use case came forward. They both had to do with the online story platform. This platform was due to be completely overhauled and new innovations were sought to enrich the new platform. Here the extra data provided by linking the GTAA to Wikidata proved useful. The first use case to sprout from the insight into the statistics manifested itself in the form of a provider of extra information for a story. It became apparent that the enriched data could both be used to display extra information on persons mentioned in the story and also to show any video content NISV might own of said person. Another use of the enriched data for the new story platform was the usage of the data as metadata to form some sort of metadata backbone for an automated relevant story system. The properties belonging to a person mentioned in a story would then be saved as metadata for that story. When two stories would have enough matching metadata they could show up as a relevant story on the bottom of the page.

5 USE CASES

Different Use cases as extracted from the interview and earlier discussions in table 5 below:

Table 5. Use cases

Name	Description
UC-1	Receiving an alert when the copyright on a person's work expires.
UC-2	Using Wikidata data to provide more information on a person when said person appears in an online story.
UC-3	Using Wikidata data as metadata to show ”Stories you might also like”.

5.1 UC-1

The GTAA mainly contains only a person's name and maybe some extra label containing some extra information to distinguish said person from other persons with the same name. Wikidata contains more data than that. Included in this data is the data of a person's date of death. When alignments are made between the GTAA and Wikidata the date of death data contained within Wikidata could be used to enrich the GTAA and enable us to get an alert on the next January, 70 years after a person's death. Analytic research into subset A showed that 35.21% of this subset have a date of death. The full results of this research can be seen in table 3 shown earlier.

Using these statistics, we can determine that we should at maximum expect a return of 3,644 people for the alert. However, we determined that further focus was needed to ensure that we would only get relevant persons in our alert. This focus had to be adjusted to the views of NISV. Meaning, there should be focus on people that worked in the fields of television, movies, radio, theater and music. To ensure that only alerts for these kind of people are received, statistics on all the different occupations of subset A were generated. A sample of the results is shown in table 6. The full list can be found in the online file sharing platform[5]. In total we found that there were 798 different occupations. We took all the occupations that had a higher occurrence rate than 0.1% as this provided a large enough sample size. Also, as people can have multiple occupations listed, the relevant occupations with a lower occurrence rate than 0.1% would still be contained within our resulting set of persons.

Table 6. Occupation distribution subset A

Occupation	Count	Percentage	Rank
politician	1,722	7.59%	1
writer	1,344	5.93%	2
actor	1,206	5.32%	3
journalist	976	4.30%	4
singer	780	3.44%	5
television presenter association	759	3.35%	6
football player	631	2.78%	7
film actor	560	2.47%	8
university teacher	538	2.37%	9
composer	516	2.28%	10
screenwriter	487	2.15%	11
presenter	485	2.14%	12
painter	433	1.91%	13
film director	394	1.74%	14
television actor	339	1.49%	15
poet	319	1.41%	16
diplomat	312	1.38%	17
lawyer	267	1.18%	18
sport cyclist	229	1.01%	19
pianist	226	1.00%	20
Other	10,162	44.80%	
Total	22,685	100.00%	

In this large list of occupation occurrence rate we picked all the occupations that had something to do with the fields of television, movies, radio, theater and music. The set of persons paired with the relevant occupations was then returned. This returned a set of 1,626 persons that had a link to the GTAA, a date of death and were relevant for NISV. With the obtained set this data can now be used to receive alerts on copyright expiration. NISV uses the Google package for their daily activities. This means that the best option for setting up such an alert system would be in a platform like Google calendar. The data was shaped to adhere to the import format of

[5] https://figshare.com/articles/Statistics/6859595

Google Calendar. After the reshaped data fit the format laid out by Google we were able to generate a view as presented in figure 1 and 2. For this example we look at January 1st 2030.

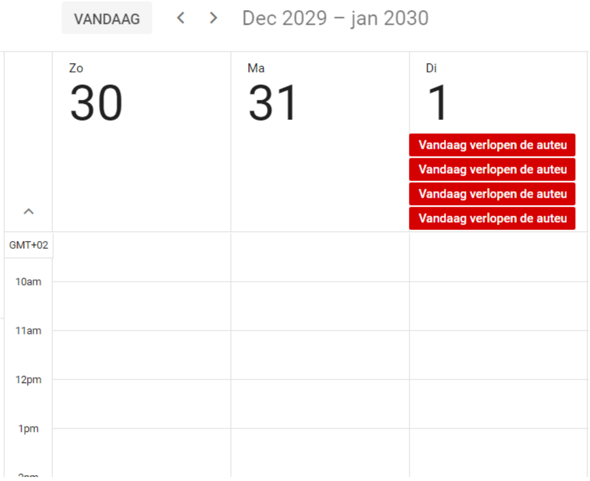
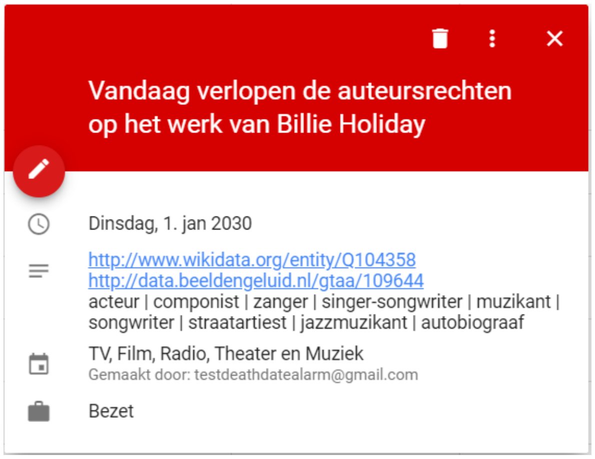

VANDAAG
Dec 2029 – jan 2030
Zo
30
Ma
31
Di
1
Vandaag verlopen de auteu
Vandaag verlopen de auteu
Vandaag verlopen de auteu
Vandaag verlopen de auteu
GMT+02
10am
11am
12pm
1pm

Fig. 1. Example UC-1



Vandaag verlopen de auteursrechten op het werk van Billie Holiday

Dinsdag, 1. jan 2030

http://www.wikidata.org/entity/Q104358
http://data.beeldengeluid.nl/gtaa/109644
acteur | componist | zanger | singer-songwriter | muzikant | songwriter | straatartiest | jazzmuzikant | autobiograaf

TV, Film, Radio, Theater en Muziek
Gemaakt door: testdeathdatealarm@gmail.com

Bezet

Fig. 2. Example UC-1

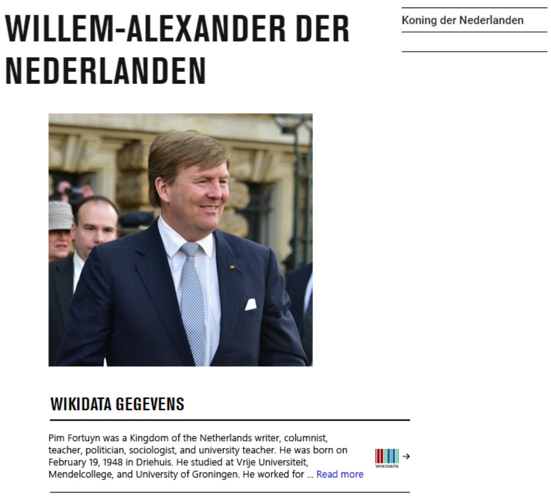
This use case was worked out as a functional prototype in Google calender.

5.2 UC-2

The second use case that came forward during the interviews was the ability to provide extra information on persons mentioned in the new story platform. A version of the story platform is currently already active. This platform is still to be completely redesigned and more resources will in the future be allocated towards this platform. When using the data available in Wikidata it enables one to quickly get a small overview of the persons mentioned in the story. Currently, the author of the story is already displayed on the right. With the use of the data from Wikidata we could expand that view to also show the persons mentioned in the story on the bottom of the page.

When someone clicks on one of the persons on the bottom of the story he/she is redirected to a separate webpage containing extra information on said person. Just like the author page that already exists. It also shows other stories the person is mentioned in. The difference between these two pages is that the information on the persons mentioned in the stories is fetched and generated from Wikidata. Wikidata already has an automated description generator that works with the properties of a person. An example of this automated service for the randomly chosen entry on Pim Fortuyn (Q311697) is as follows: ”Pim Fortuyn was a Kingdom of the Netherlands writer, columnist, teacher, politician, sociologist, and university teacher. He was born on February 19, 1948 in Driehuis. He studied at Vrije Universiteit, Mendelcollege, and University of Groningen. He worked for Maastricht University, for Erasmus University Rotterdam, and for University of Groningen. He died of ballistic trauma by Volkert van der Graaf on May 6, 2002 in Hilversum. He was buried at San Giorgio della Richinvelda. ; Dutch politician”. For this example, another well known Dutch person was chosen as Willem-Alexander is not actually linked to the GTAA yet. An example of how the page with extra information on Willem-Alexander with the automatic generated description of Pim Fortuyn (Q311697) can be seen in Figure 3 on the next page.

This provides the reader with extra information on said person. A second use of this page could be to show existing relevant video footage about the person



WILLEM-ALEXANDER DER NEDERLANDEN

Koning der Nederlanden

WIKIDATA GEGEVENS

Pim Fortuyn was a Kingdom of the Netherlands writer, columnist, teacher, politician, sociologist, and university teacher. He was born on February 19, 1948 in Driehuis. He studied at Vrije Universiteit, Mendelcollege, and University of Groningen. He worked for ... Read more

Fig. 3. Example UC-2

available for the public. This could be shown below the automatically generated description.

In conclusion, this usage of Wikidata data retrieval can be used by the author to provide extra information on the persons mentioned in the story. The authors themselves can decide if they feel it necessary to display this extra information. Also the authors themselves can decide for which persons in story an extra page could be created. The choice for this extra functionality at the bottom of the page is based on the fact that the only readers interested in the extra information/video material on the persons mentioned in the stories would be the ones that actually finish the whole story. It also causes less distraction at the start of the story. The reason this would likely not completely work as a fully automatic tool is that not all the persons mentioned in the story are as relevant to the story or in general. This use case was worked out as concept with illustrations to support it.

5.3 UC-3

A third use case for the data from Wikidata for the story platform is using the available data as a metadata backbone. When persons are mentioned in the story their data available on Wikidata could be used and stored as metadata for the story. When doing this one could generate an automated ”Stories you might also like” on the bottom of a story. When looking at an example from the story ”Hoe Cronkite de kijk op de Vietnamoorlog veranderde”[6]. We see that Walter Cronkite, William Westmoreland, Lyndon B. Johnson and Creighton Abrams are mentioned in the story. All properties of these persons can be retrieved from Wikidata. All the properties of Walter Cronkite (Q31073) are found in table 7 on the next page.

If we look at these properties it can be concluded that not every property is useful for use as metadata. If properties like ”Place of death” would be used as metadata it would not return relevant stories. This is attributed to the fact that the location of one’s death does not necessarily define a person. The ”Place of death” property is not alone in this. On the same note ”Birth name”, ”Given name”, ”Manner of death” and ”Cause of death” are less relevant. To tackle this problem, we would have to rate all possible properties on a scale of ”Not relevant” to ”Relevant”. Secondly, all properties are also bound to a category as this could also help with determining relevance. An overview of all these properties is shown in Appendix A.

This leaves us with a significantly smaller list of properties. When using this data on relevancy for the example case of ”Hoe Cronkite de kijk op de vietnamoorlog veranderde” we see that the relevant properties for Walter Cronkite (Q31073) would be: sex or gender, country of citizenship, award received, date of birth, child, religion and date of death. This data could be linked to metadata generated from other stories to support the author of the story in making a decision on the relevant story selection. For this particular example of Walter Cronkite (Q31073) it would trigger on other stories about people with similar properties. This in itself would not in any way guarantee a truly relevant story. For example: someone’s country of citizenship is relevant, but does not have to make a story related to another story. The best way to handle this is to determine some sort of threshold for matching properties. If a threshold is set for a percentage of matching properties for the metadata of stories we can get a more relevant view. The author of the story can then look at

[6] https://www.beeldengeluid.nl/verhalen/hoe-cronkite-de-kijk-op-de-vietnamoorlog-veranderde

Table 7. All properties of Walter Cronkite (Q31073)

Instance of	Human	**Date of birth**	4 November 1916	**Cause of death**	Stroke
Sex or gender	Male	**Place of birth**	St. Joseph	**Child**	Kathy Cronkite
Country of citizenship	United States of America	**Date of death**	17 July 2009	**Languages written, spoken or signed**	English
Birth name	Walter Leland Cronkite, Jr. (English)	**Place of death**	New York City	**Occupation**	Journalist, News presenter, Blogger
Given name	Walter	**Manner of death**	Natural causes	**Member of**	American Academy of Arts and Sciences, American Philosophical Society
Award received	Presidential Medal of Freedom	**Educated at**	University of Texas at Austin	**Religion**	Episcopal Church

the automatically generated relevant story and determine if this automatically generated match is viable as a related story.

As a second part of this use case is the different categories in which the different properties are divided could be linked to the user profile. This can be done in a sense that if a person is classified in the category ”History” the properties belonging to that category could be weighed higher than the others. This information could then be used to better generate relevant stories for persons interested in those categories.

In conclusion, this tool can be used by the author to give him insight in automatically generated relevant stories. The predictions would be too random to actually use as a completely automatic tool. With this tool the authors themselves can decide if one of the generated relevant story is one they themselves actually overlooked. This use case was worked out as a conceptual tool.

5.4 Patterns in use cases

When looking at the way the use cases were set up we see that the most important thing to do is collecting all the possible relevant data. This allowed us to give a clear insight in what the dataset actually contains. UC-2 and UC-3 were impossible to set up without first allowing internal users to get an overview of the dataset.

That being said we delve deeper into what allowed us to set up these use cases. In essence, all the use cases start with a focus on a specific aspect of the properties. UC-1 came forth from a focus on the specific property of ”Date of death”. UC-2 came forward when focusing on properties with a high occurrence rate. UC-3 was made possible by focusing on a very specific set of properties. The pattern we can derive from this is that one can focus on a specific set of properties to find a new use case. We also see the differences in the setting up of use cases this way. The similarities are simple to spot; the extra data brings forth new functionality.

If we would look further into this process of specific property targeting, we would be able to actually come up with extra use cases. For example, if we would once again focus on properties with a high occurrence rate we would be able to set up an extra use case in which we could divide the persons in certain categories. Those categories would be based upon the properties with an exceptional high occurrence rate.

6 EVALUATION

After the use cases were set up and tested, an evaluation took place. This evaluation consisted of two evaluation rounds. In the first round internal users of the department of Knowledge & Innovation were shown the worked out use cases in a worked out document. They were asked if the worked out use cases were what they had envisioned during the interview stages. They were also asked if they would make any improvements or changes to the use case as stated in the documentation. After the proposed changes were made to the use cases there was a second round of evaluation. The final document with the updated use cases was discussed in a meeting between internal users at the department of Knowledge & Innovation. After this meeting informal conversations were conducted with the people involved. These conversations allowed us to get in insight into the projected usefulness of the use cases.

After the first phase we found a unanimous proposed change for UC-1. This proposed change was to further focus on people with a media background. In the earlier worked out version all persons that have a date of death were included in the worked out prototype. According to the evaluation this caused a lot of unnecessary alerts. After receiving this review changes were made so that only persons with a media background were included in the alarm. There were also proposed changes to UC-2. For UC-2 the proposed change was to include possible archive material on the extra information page for a person mentioned in a story. It was also proposed to move extra information button from the top right to the bottom of the page as it could prove too distracting. UC-3 did not receive any proposed changes after the first phase.

After the second phase of evaluations informal conversations between different internal users took place. The updated UC-1 was evaluated as a useful tool for keeping track of copyright expirations. Further alignments between the GTAA and Wikidata are still needed to tackle the problem of incompleteness of the data, but the foundation of the use case was well received. UC-2 and UC-3 were evaluated as interesting assets for the new story platform. Their actual impact and usefulness was hard to be determined by the internal users as they were both still conceptual use cases.

7 DISCUSSION & FUTURE WORK

7.1 Discussion

This research also had some limitations. First of all, we were limited in our use of the SPARQL endpoint tool provided by Wikidata. When doing the statistical analysis, we used this tool to retrieve all our information from Wikidata. This tool had its limitations as too complex or too large queries would result in the service timing out. Something that could have been done better in this research was better SPARQL query optimization. Due to time pressure we were not able to divert more attention into this matter. Another solution to the timing out problem would be to run a Wikidata RDF dump in a local virtual environment. This was however also something we had to abandon because of time pressure. Especially the second option can prove useful for future research into a similar subject as it would enable the researchers to have full control over the dataset.

When performing similar case studies one thing that should be considered is the combination of a data driven approach and a user driven approach. In this research the combination of said approaches returned good results in the form of use cases. It also brought forth a framework for said use cases in the form of an analysis on the patterns. Without the combination of both approaches the use case would not be retrievable in their current form.

7.2 Future work

To ensure a better future for the set up use cases, the production of alignments between the GTAA and Wikidata needs to continue. As mentioned before only 8.4% of the possible alignments have been made. Therefore, the use cases as of now cannot be used as a reliable tool for getting the needed information. This is all attributed to the fact that the dataset is incomplete. Further coordination between the volunteers of Wikidata and the internal users at NISV is needed.

As far as the broader part of this research is concerned, this use case study proved an excellent starting point for further research into the practical use of linked datasets. A lot of research has already been conducted in the different online knowledge bases (e.g. Wikidata, DBpedia) and their usefulness. However, this research shed some new light on the usefulness and use cases when linking big datasets to those online knowledge bases. The method of starting with a data driven approach and continuing using a user driven approach showed how a mix of the two approaches can be used to effectively determine the use cases in similar case studies.

In this sense, this research hopes to spark future research in other specific use cases for the usefulness of linked data.

8 CONCLUSION

This research aimed to find answers to the questions of what use cases can be formed by an enriched GTAA and the differences in wishes between internal and external users are. By looking deep into the data both contained in the GTAA and Wikidata it found some meaningful differences but also similarities. What can be concluded from the statistical analysis is that the subset A is representative of dataset B in a sense of gender distribution and some other generic properties. The meaningful differences lie in the more specific properties like ”Award received”. Here it became apparent that subset A was much more media oriented than subset B.

This research also showed that when giving users insight in all the raw data and its properties, new ideas about the usefulness of said data sparks. In this case study this manifested itself in the form of use cases. We conclude that focus on different properties within the data lead to completely different and independent use cases. The use cases that came forward in this research were connected to a specific focused group of properties. For UC-1 the focus lied on the ”date of death” property. For UC-2 the focus lied on properties with a high occurrence rate. For UC-3 the focus lied on very specific properties that were handpicked for relevancy.

In the evaluation the usefulness of all three use cases came forward. UC-1 in as of now a working tool for the internal users. They can themselves opt to use it to better manage copyright expiration cases. UC-2 and UC-3 are still conceptual use case, but were evaluated to be good in concept. UC-3 is thought to be so interesting as a concept by internal users of the department Knowledge & Innovation that a future project to calculate its usefulness in exact values is now being set up.

For an answer to the second research question we will have to look at the group of external users. Those external users are defined as external researchers that make use of the facilities at NISV to further their own research. A notable difference between the internal and external users lie in their goals, tools and ways of work. The internal users look at the usefulness for an enriched GTAA for the development of their own projects whereas the external users would use an enriched GTAA to only further their research. External users also use tools that internal users do not. They mainly use the CLARIAH Mediasuite. In this tool the GTAA is an optional component. Internal users use the GTAA in a different way and also use different tools in managing it. These tools (Like SKOS) are not open to external researchers. There are differences in the wishes of the two groups of users. Internal users want to mainly use an enriched GTAA to set up further project or create new tools. Whereas the external users mainly want to use an enriched GTAA to receive as much extra information as possible as well as add some functionality to their search behaviour. The similarities in the wishes lie in the wish for extra data and thus extra functionality to certain aspects of their workload.

REFERENCES

[1] Vrandečić & Krötzsch, M. (2014). Wikidata: a free collaborative knowledgebase. Communications of the ACM, 57(10), 78-85.

[2] Erxleben, F., Günther, M., Krötzsch, M., Mendez, J., & Vrandečić, D. (2014, October). Introducing Wikidata to the linked data web. In International Semantic Web Conference (pp. 50-65). Springer, Cham.

[3] Färber, M., Bartscherer, F., Menne, C., & Rettinger, A. (2016). Linked data quality of dbpedia, freebase, opencyc, wikidata, and yago. Semantic Web, (Preprint), 1-53.

[4] Thornton, K., Cochrane, E., Ledoux, T., Caron, B., & Wilson, C. (2017). Modeling the Domain of Digital Preservation in Wikidata.

[5] Debevere, P., Van Deursen, D., Mannens, E., Van de Walle, R., Braeckman, K., & De Sutter, R. (2011). Linking thesauri to the linked open data cloud for improved media retrieval. In 12th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS-2011).

[6] Tordai, A., Omelayenko, B., & Schreiber, G. (2007, October). Thesaurus and metadata alignment for a semantic e-culture application. In Proceedings of the 4th international conference on Knowledge capture (pp. 199-200). ACM.

[7] Kobilarov, G., Scott, T., Raimond, Y., Oliver, S., Sizemore, C., Smethurst, M., ... & Lee, R. (2009, May). Media meets semantic webfi?!how the bbc uses dbpedia and linked data to make connections. In European Semantic Web Conference(pp. 723-737). Springer, Berlin, Heidelberg.

[8] Caracciolo, C., Stellato, A., Rajbahndari, S., Morshed, A., Johannsen, G., Jaques, Y., & Keizer, J. (2012). Thesaurus maintenance, alignment and publication as linked data: the AGROVOC use case. International Journal of Metadata, Semantics and Ontologies, 7(1), 65-75.

[9] de Len, A., Saquicela, V., Vilches, L. M., Villazn-Terrazas, B., Priyatna, F., & Corcho, O. (2010, September). Geographical linked data: a Spanish use case. In Proceedings of the 6th International Conference on Semantic Systems (p. 36). ACM.

[10] Neubert, J. (2009). Bringing the" Thesaurus for Economics" on to the Web of Linked Data. LDOW, 25964.

[11] Klein, M., Gupta, H., Rai, V., Konieczny, P., & Zhu, H. (2016, August). Monitoring the Gender Gap with Wikidata Human Gender Indicators. In Proceedings of the 12th International Symposium on Open Collaboration (p. 16). ACM.

A APPENDIX A - ALL PROPERTIES WIKIDATA

Property	**Data type**	**Category**	**Relevancy**
sex or gender	item	Generic	Relevant
date of birth	Point in time	Generic	Relevant
birthday	item	Generic	Relevant
place of birth	item	Generic	Relevant
birth name	Monolingual text	N/A	Not relevant
date of death	Point in time	Generic	Could be relevant
place of death	item	N/A	Not relevant
cause of death	item	N/A	Not relevant
manner of death	item	N/A	Not relevant
killed by	item	N/A	Not relevant
place of burial	item	N/A	Not relevant
image of grave	Commons media file	N/A	Not relevant
name in native language	Monolingual text	N/A	Not relevant
ancestral home	item	N/A	Not relevant
member of	item	Generic	Relevant
ethnic group	item	Generic	Relevant
native language	item	Generic	Relevant
country of citizenship	item	Generic	Relevant
educated at	item	Generic	Relevant
occupation	item	Generic	Relevant
field of work	item	Generic	Relevant
notable work	item	Journalism, Music, Documentary, Film, Television, Entertainment, Webvideo, Games, Art & culture and History	Could be relevant
employer	item	Generic	Could be relevant
award received	item	Generic	Relevant
position held	item	Journalism, History and Organization	Relevant
member of political party	item	Journalism and History	Relevant
residence	item	Generic	Could be relevant
official residence	item	Generic	Could be relevant
religion	item	Generic	Relevant
sexual orientation	item	Generic	Could be relevant
coat of arms image	Commons media file	N/A	Not relevant
coat of arms	item	N/A	Not relevant
signature	Commons media file	N/A	Not relevant
doctoral advisor	item	N/A	Not relevant

Property	**Data type**	**Category**	**Relevancy**
doctoral student	item	N/A	Not relevant
student of	item	N/A	Not relevant
student	item	N/A	Not relevant
canonization status	item	N/A	Not relevant
voice type	item	N/A	Not relevant
shooting handedness	item	N/A	Not relevant
astronaut mission	item	N/A	Not relevant
dan/kyu rank	item	N/A	Not relevant
Eight Banner register	item	N/A	Not relevant
family	item	Generic	Relevant
noble title	item	N/A	Not relevant
honorific prefix	item	N/A	Not relevant
academic degree	item	Generic	Relevant
handedness	item	N/A	Not relevant
website account on	item	N/A	Not relevant
honorific suffix	item	N/A	Not relevant
family name	item	Generic	Relevant
given name	item	N/A	Not relevant
pseudonym	String	N/A	Not relevant
feast day	item	N/A	Not relevant
audio recording of the subject's spoken voice	Commons media file	N/A	Not relevant
manager/director	item	Journalism, Documentary, Film, Television, Entertainment, Webvideo, History and media-use	Relevant
filmography	item	Journalism, Documentary, Film, Television, Entertainment, Webvideo, History and media-use	Relevant
instrument	item	Music and Radio	Relevant
eye color	item	N/A	Not relevant
participant of	item	Sport and History	Relevant
convicted of	item	N/A	Not relevant
languages spoken, written or signed	item	Generic	Could be relevant
affiliation	item	Organization	Relevant
has pet	item	N/A	Not relevant
Commons Creator page	String	N/A	Not relevant
Project Gutenberg author ID	External identifier	N/A	Not relevant
father	item	Generic	Relevant
mother	item	Generic	Relevant
sibling	item	Generic	Relevant
spouse	item	Generic	Relevant

Property	**Data type**	**Category**	**Relevancy**
partner	item	Generic	Relevant
child	item	Generic	Relevant
stepparent	item	Generic	Relevant
relative	item	Generic	Relevant
godparent	item	Generic	Could be relevant
number of children	quantity	N/A	Not relevant
member of sports team	item	Sport	Relevant
Position played on team / specialty	item	Sport	Relevant
Doubles record	String	N/A	Not relevant
Singles record	String	N/A	Not relevant
ranking	Quantity	N/A	Not relevant
Country for sport	Item	Sport	Could be relevant
Swedish Olympic Committee athlete ID	External identifier	N/A	Not relevant
Military branch	item	History	Could be relevant
Military rank	item	History	Could be relevant
Commander of	item	History	Could be relevant
conflict	item	History	Could be relevant