Assessing missing data for alternative music groups

Adrian Rodriguez Grillo - 6193748

March 1, 2019

1 Introduction

Alternative or indie music was born to oppose the rules, characteristics and ways popular melodies were made, the term emerged during the 1980s and became popular in the 1990s. In the origin, this style referred to underground bands that sound different to the mainstream rock that was prevalent in the epoch, however, nowadays the definition has become broader and affect all the existent musical genres, like pop, hip-hop, electronic, etc.

Although the sound of the songs is the main characteristic that differs from the parameters of the mainstream styles, in general, alternative bands are also identified for having different ideas and ways of how the music is produced. Most of these groups are out of the commercial grid and do not have much influence in scene, releasing their albums under independent records without much or any publicity and counting with his fans to make themselves known. Furthermore, indie groups are generally more compromised with the social context and their origin region, something that will be taken into consideration in this project.

Regarding the distribution of the music, internet has become the platform that provides the greatest revenues [15] and present a great opportunity to this type of bands to make their music known. The availability of complete and correct information on this platform provides a better position in searches, allowing the access to a broader audience. Public domain data platforms like Wikipedia [8], Wikidata [7], DBpedia [2] or, more specialized ones, like MusicBrainz [6] facilitates the publication of this information, however, the lack of resources of these bands leads to a situation where the data differs between websites or, directly, is missing.

2 Problem and methodology

A great quantity of information about alternative and underground bands exist on the internet, most of them aggregated by the fans but also by private companies that are dedicated to the music business, like Spotify [16] or Last.fm [4]. However, the data could vary in a great quantity depending of the platform visited or the search provider used, leading to inconsistencies.

In general, these private pages contain more accurate and correct data but that is not always the case, moreover, the data is not freely available and is usually being subject to some conditions. On the contrary, is easy to find these problems in open data portals with situations where the genres of a band differ between sites or where some information is missing in some but not others.

In order to solve this issue, this project will study the possibility of completing the data of the open data portals using the information contained in them, this is using interlinking. Moreover, the possibility of using external resources, like music services APIs, to complete the information will be also considered. The main objective in this project will be to complete the genre and the geographical information of the groups, however, other data will be also considered.

To accomplish the objective, the MusicBrainz dataset will be used as main source of music information, generating a knowledge graph with the data contained in the database. This process will be done manually due to the lack of an updated dump of the data in RDF format. The Music Ontology [1] will be used as the vocabulary in order to make the data more generic and easily expandable.

The Wikidata and DBpedia knowledge graphs will be also used and linked to the data using the tools available, like the Silk framework or LIMES. Apart from the linking, the objective will be the join of the information available, filling the missing data in the correspondent sources and, generally, completing it. The GeoNames [9] dataset is also intended to be used to give importance to the geographical information of the bands and proportionate another way of relation that could improve the discovery possibilities.

The use of external and not structured data to complete the information of the graph will be also studied. As music data is being handled, the focus will be the use of APIs of music services to complete the data in an automatic way, however, the possibility of using more unstructured sources like Wikipedia or search engines will be reviewed.

3 Milestones and deliverables

The main outcome of this project will be the publication of a knowledge graph that will use the MusicBrainz data as base and the Music Ontology as the main dictionary. This knowledge graph will be linked to the data contained in Wikidata, DBpedia and GeoNames.

Additionally, a study about how is possible to tackle the completion of the data using the linked information will be assessed. This is, how to update the content of a specific source with the knowledge contained in the linked entities, aiming for a release of a tool that performs this kind of completion.

Finally, the use of external and unstructured data to complete the information contained in knowledge graphs will be reviewed, focusing specifically in music information and the use of dedicated webservices. The implementation of an automatic tool will be contemplated as well as a review of existent methods that perform this task.

4 Related work, novelty and significance

There have been previous projects that had used the MusicBrainz dataset to generate a knowledge graph in order to improve the connectivity with the linked data initiative, however they have been discontinued because of the lack of resources. Nevertheless, some official information exists about the process of generating a graph [12] and, also, a non-official SPARQL endpoint [5].

Linking knowledge graphs have been a widely covered topic and one of the main purposes of the semantic web. Tools like LIMES [13] and the Silk framework [18] facilitates the task of interlinking different knowledge graphs and in [10] a review all the principles related with the task can be found. Additionally, there have been some efforts to use external and unstructured information to expand and improve existent knowledge graphs [17, 11, 3].

Although in [14] the possibility of using interlinked data is contemplated to complete the knowledge graph information, the mentioned approaches effort goes into predict either missing entities, missing types for entities, and/or missing relations that hold between entities. Therefore, there is not much work that covers the problem of completing and filling the information of existent entities with data that links to them.

Therefore, this project will build a music specific knowledge graph using techniques and methods that have been successful applied in the past to, later, improve and complete the existent data using a previously unexplored approach. Additionally, even though the use of external data to improve the information is not novel, using specific tools could add some value to the ecosystem.

As a passionate fan of alternative music and, specially, of small groups of my country, I think that any help this kind of groups can receive is extremely valuable for them as it increases the possibility of being discovered. Moreover, as a defender of open data and crowdsourced information I believe that any project that can help or give ideas on how to improve the quality of the data are extremely important nowadays. Specifically, if this project success it could help to improve the consistency among the different open data sources.

References

[1] Samer Abdallah, Yves Raimond, and Mark Sandler. “An Ontology-based Approach to Information Management for Music Analysis Systems”. In: vol. 1. May 2006.

[2] Sören Auer et al. “DBpedia: A Nucleus for a Web of Open Data”. In: *PROC. 6TH INT’L SEMANTIC WEB CONF.* Springer, 2007.

[3] Diego Collarana et al. “Synthesizing Knowledge Graphs from web sources with the MINTE+ framework”. In: Oct. 2018.

[4] CBS Corporation. *Last.fm.* URL: https://www.last.fm/.

[5] DBTune. *MusicBrainz D2R server.* URL: http://dbtune.org/musicbrainz/.

[6] MetaBrainz Foundation. *MusicBrainz.* URL: https://musicbrainz.org/ (visited on 02/28/2019).

[7] Wikimedia Foundation. *Wikidata.* URL: https://www.wikidata.org (visited on 02/28/2019).

[8] Wikimedia Foundation. *Wikipedia.* URL: https://en.wikipedia.org/wiki/Wikipedia (visited on 02/28/2019).

[9] *GeoNames.* URL: https://www.geonames.org.

[10] Tom Heath and Christian Bizer. *Linked Data: Evolving the Web into a Global Data Space.* Vol. 11. Feb. 2011. DOI: 10.2200/S00334ED1V01Y201102WBE001.

[11] Heng Ji and Ralph Grishman. “Knowledge Base Population: Successful Approaches and Challenges.” In: Jan. 2011, pp. 1148–1158.

[12] *Linked data open cloud.* URL: https://lod-cloud.net/.

[13] Axel-Cyrille Ngonga Ngomo and Sören Auer. “LIMES – A Time-Efficient Approach for Large-Scale Link Discovery on the Web of Data”. In: Jan. 2011, pp. 2312–2317. DOI: 10.5591/978-1-57735-516-8/IJCAI11-385.

[14] Heiko Paulheim. “Knowledge graph refinement: A survey of approaches and evaluation methods”. In: *Semantic Web* 8 (Dec. 2016), pp. 489–508. DOI: 10.3233/SW-160218.

[15] International Federation of the Phonographic Industry. *IFPI Global Music Report 2018*. 2018. URL: `https://ifpi.org/news/IFPI-GLOBAL-MUSIC-REPORT-2018` (visited on 02/28/2019).

[16] Spotify Technology SA. *Spotify*. URL: `https://www.spotify.com`.

[17] Baoxu Shi and Tim Weninger. "Open-World Knowledge Graph Completion". In: *CoRR* abs/1711.03438 (2017). arXiv: `1711.03438`. URL: `http://arxiv.org/abs/1711.03438`.

[18] Julius Volz et al. "Discovering and maintaining links on the web of data". In: *In The Semantic Web – ISWC 2009: 8th International Semantic Web Conference*. 2009, pp. 650–665.