# ELMD: An Automatically Generated Entity Linking Gold Standard Dataset in the Music Domain

**Sergio Oramas*, Luis Espinosa-Anke†, Mohamed Sordo**, Horacio Saggion†, Xavier Serra***

* Music Technology Group, † Tractament Automàtic del Llenguatge Natural, ** Center for Computational Science
†* DTIC - Universitat Pompeu Fabra, Barcelona, Spain ** University of Miami, Coral Gables, FL, USA
{sergio.oramas, luis.espinosa, horacio.saggion, xavier.serra}@upf.edu, mohamed.sordo@gmail.com

**Abstract**

In this paper we present a gold standard dataset for Entity Linking (EL) in the Music Domain. It contains thousands of musical named entities such as Artist, Song or Record Label, which have been automatically annotated on a set of artist biographies coming from the Music website and social network LAST.FM. The annotation process relies on the analysis of the hyperlinks present in the source texts and in a voting-based algorithm for EL, which considers, for each entity mention in text, the degree of agreement across three state-of-the-art EL systems. Manual evaluation shows that EL Precision is at least 94%, and due to its tunable nature, it is possible to derive annotations favouring higher Precision or Recall, at will. We make available the annotated dataset along with evaluation data and the code.

**Keywords:** Entity Linking, Music Information Retrieval, Language Resources

## 1. Introduction

When we refer to the Music Domain in a Natural Language Processing (NLP) context we refer to Music product reviews such as Albums or Songs, Music-related biographies or even song lyrics. While these are valuable resources in NLP for tasks like Sentiment Analysis, Music Information Retrieval (MIR), however, has barely exploited the information and knowledge that can be extracted from textual data. This opens up a vibrant area of research where MIR tasks may benefit dramatically from mining textual data (Sordo et al., 2015). For example, Music Recommendation is increasingly leveraging NLP in order to go beyond shallow Music item features and shared listening habits (Ostuni et al., 2015). However, the intersection between NLP and MIR is vastly underdeveloped, with few available evaluation frameworks, Music-specific corpora and *ad-hoc* tools.

An interesting common ground for exploiting the intersection between NLP and MIR is the automatic generation of Music Knowledge Bases (KBs). These may be further exploited for computing Artist Similarity (Oramas et al., 2015a) or Music Recommendation (Sordo et al., 2015). While it is not the purpose of this paper to provide an overview of methods for KB learning, let us mention that generally the approach starts by processing large corpora in order to obtain an annotation for Music entity mentions in text, either simply as Music types (e.g. tagging 'Yellow Submarine' as `Song`) or performing Entity Linking (EL), e.g. tagging 'Yellow Submarine' as `dbpedia.org/page/Yellow_Submarine_(song)`. However, this is not a trivial task as mentions to Music entities show language and register idiosyncrasies (Tata and Di Eugenio, 2010; Gruhl et al., 2009), and therefore a certain degree of tailoring is required in order to account for them. Let us consider multiword Music entities, which usually are those who pose greatest challenges for EL. As (Tata and Di Eugenio, 2010) point out, they are difficult to discover because they may not be restricted to a single Noun Phrase or may be abbreviated (by means of acronyms, dropping entire words or even full rephrasing). Additionally, a specific trait of Music texts is the fact that one song may have many covers by many different artists and, according to our evaluation, it may be difficult even for a human to identify what *version* of the song the writer is referring to. Furthermore, availability of EL testbeds in general (Usbeck et al., 2015), and in the Music domain in particular (Gruhl et al., 2009), is scarce, making it very difficult to evaluate novel systems and approaches. Hence, it is difficult to know how well a certain method, which may work well for generic texts, will perform on Music data.

Despite the current context of scarcity of both EL systems and evaluation benchmarks in the Music domain, there are some exceptional cases in which these issues were addressed, such as: (1) Detecting Music entities (e.g. songs or bands) on informal text (Gruhl et al., 2009); (2) Applying Hidden Markov Models for discovering Music entity mentions in Chinese corpora (Zhang et al., 2009); or (3) Recognizing musical entities in the context of a relation extraction pipeline (Oramas et al., 2015b).

We argue that the problem of precision in detecting musical entities may be tackled by leveraging a combination of several generic EL off-the-shelf systems. Simply put, we hypothesize that if two or more generic systems annotate with the same URI an entity mention, the probability of this annotation to be correct increases. To the best of our knowledge, very little effort has been put in exploiting this *agreement* feature. One of the reasons may be that, as of now, most EL systems *speak their own language*, partially due to the fact that each of them points back to different KBs, and hence their output is heterogeneous and cannot be directly compared, let alone combine. This has motivated research towards unification frameworks for evaluation of EL. For instance, (Cornolti et al., 2013) put forward a benchmarking framework for comparing EL systems. Moreover, (Rizzo and Erp, 2014) describe a system aimed at combining the output of the different NER sys-

tems. Finally (Usbeck et al., 2015) present GERBIL, an evaluation framework for semantic EL based on (Cornolti et al., 2013).

In this paper we aim to provide a twofold answer to the challenges described above, and bridge the gap between the Music domain and EL. Specifically, we present ELMD, an automatically constructed corpus where named entities are classified as any of four predefined *musical categories*, namely SONG, ALBUM, ARTIST, and RECORD LABEL, by leveraging the hyperlinks present in a set of artist biographies. Then, we further enrich ELMD by performing EL and automatically annotating a large portion of the entities with their DBPEDIA URI. The source data used in this paper comes from the music website and social network Last.fm[1]. To the best of our knowledge, this is the first attempt to provide an annotated large-scale corpus of linked entities in the music domain.

The final resource amounts to 47,254 sentences, in which 92,930 entities are categorized into the aforementioned *musical categories*, and 64% of them are disambiguated and linked to DBpedia. We achieve a Precision score of 97% in the most restrictive setting, in which our approach manages to annotate more than 31,000 entities.

In the remainder of this paper, we first introduce ELVIS (Entity Linking Voting and Integration System), our EL integration and agreement approach. Then, we describe the text corpus we compiled from the LAST.FM website and how it is combined with ELVIS. In the next step, the obtained dataset is evaluated. Finally, we describe the resulting output of our system: The ELMD dataset.

## 2. ELVIS

In this section we describe ELVIS, the generic integration framework for Entity Linking, which is leveraged for the construction of ELMD. First, we describe our Entity Linking research problem and provide an intuition on how this may be surmounted via an agreement scheme. Then, we provide details on the main modules integrating ELVIS, highlighting the possible cases of agreement and disagreement over the EL systems that are integrated in our framework.

### 2.1. Argumentum ad Populum in EL

Our method relies on the *argumentum ad populum* intuition, i.e. if two or more different EL systems perform the same prediction in linking a named entity mention to its entry in a reference KB, the more likely this prediction is to be correct. We put this intuition into practice by combining the output of three well-known systems, namely DBpedia Spotlight (Mendes et al., 2011) , Tagme (Ferragina and Scaiella, 2012) and Babelfy (Moro et al., 2014), whose agreement (or disagreement) when disambiguating an in-text entity mention is taken as an agreement-driven *confidence score*. These specific tools were chosen for being considered state-of-the-art EL systems and for being well known in the NLP community. However, ELVIS can easily incorporate any additional system. We also selected these tools because entities identified by all of them can be easily referenced to DBpedia URIs. Although there are other knowledge bases (e.g. MusicBrainz) with substantially more musical entities than DBpedia, to the best of our knowledge, there is no EL tool that works with these domain specific knowledge bases.

Let us briefly describe each of the selected EL systems:

- **DBpedia Spotlight** (Mendes et al., 2011) is a system for automatically annotating text documents with DBpedia URIs, finding and disambiguating natural language mentions of DBpedia resources. DBpedia Spotlight is shared as open source and deployed as a Web service freely available for public use[2]. DBpedia Spotlight gives as a result the DBpedia URI, start and end char positions, the value of the rdf:type property, and a confidence score for for each prediction.

- **TagMe** (Ferragina and Scaiella, 2012) is an EL system that matches terms with Wikipedia link texts and disambiguates them using the in-link graph and the Wikipedia page dataset. Then, it performs a pruning process by looking at the entity context. TagMe is available as a web service[3], and provides the Wikipedia page id, Wikipedia categories, and a confidence score.

- **Babelfy** (Moro et al., 2014) is an EL and Word Sense Disambiguation based on non-strict identification of candidate meanings (i.e. not necessarily exact string matching), together with a graph based algorithm that traverses the BabelNet graph and selects the most appropriate semantic interpretation for each candidate. Babelfy is available as a web service[4]. Its output is based on the corresponding BabelNet synset of the disambiguated mention. If the synset references to a Wikipedia page, it returns the Wikipedia URL, the DBpedia URI, as well as Wikipedia categories.

While these tools have proven highly competitive on their own, in this paper we explore the gain in performance obtained by combining them together, and apply global agreement-driven decisions on the LAST.FM corpus.

### 2.2. 'Translating' EL Formats

In order to have each EL system *speak the same language* for measuring agreement in their predictions, output homogenization is required. This is not a trivial task, as each EL approach may be based on a different reference KB, the offsets may be computed differently, and so on. For instance, DBpedia Spotlight links entity mentions via DBpedia URIs, whereas Tagme provides Wikipedia page IDs, and Babelfy disambiguates against BabelNet (Navigli and Ponzetto, 2012) and its corresponding BabelNet synsets. We attempt to surmount this heterogeneity as follows: First, we retrieve DBpedia URIs of every named entity. There are some considerations to be taken into account, however: (1) Character encoding differs from system to system, which

[1] http://www.last.fm

[2] https://github.com/dbpedia-spotlight

[3] http://tagme.di.unipi.it/

[4] http://babelfy.org/guide

we address by converting the character encoding of the retrieved URI to UTF-8; (2) Several URIs may refer to the same DBpedia resource. We solve this specific issue thanks to the transitive redirections provided by DBpedia. If a URI has a transitive redirection, it is replaced by the redirected URI. (3) Note that, in the case of Tagme, only Wikipedia page IDs are provided, which we can straightforwardly exploit to map entity mentions to their DBpedia equivalent. Finally, and after surmounting compatibility issues among systems, we retrieve DBpedia types (`rdf:type property`) for all entities. This *type* information is further used in the creation of ELMD.

After successfully providing a process which harmonizes the output of EL systems, it is possible to compute the degree of agreement among them, which will become our system's confidence score. We define the following set of *agreement heuristics* to set such score for each linking prediction (an overview of the workflow of ELVIS is provided in Figure 1).

- **Full Agreement** $(++)$ When all systems detect an entity with the same URI and offset.
- **Partial Agreement** $(+)$ When more than one but less than all systems detect an entity with the same URI and offset. Outliers (i.e. systems performing a different prediction) may detect a different entity or may not detect anything.
- **Singleton Decision** $(-)$ When only one system detects an entity for a given text offset.
- **Disagreement** $(--)$ When more than one system performs a linking over the same text offset, but all of their predictions are different.

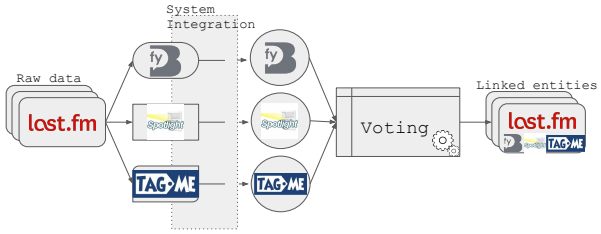


Figure 1: ELVIS Workflow

## 3. From LAST.FM to ELMD

In what follows, we describe the original data gathered from LAST.FM, and the process to apply the integration framework described in Section 2., in order to construct a highly precise benchmarking dataset for EL in the Music domain.

In LAST.FM, users may add relevant biographical details to any artist's main page in the form of a *wiki*. These edits are regularly moderated. Furthermore, artist biographies are often enriched with hyperlinks to other LAST.FM Artist, Album, Song and Record Label pages, similarly as with Wikipedia hyperlinks. Our purpose is to leverage this meta-information to automatically construct a dataset of Music-specific annotated named entities.

We crawled artist biographies from LAST.FM in March 2015, and gathered 13,000 artist biographies, which comprise 47,254 sentences with at least one hyperlink, amounting to a total of 92,930 links. These may be broken down as follows: (1) 64,873 hyperlinks referencing Artist pages; (2) 16,302 to Albums; (3) 8,275 to Song pages; and finally (4) 3,480 hyperlinks referencing Record Labels. This *type* information is extracted thanks to the structure of each link's URL, as it includes in its path the category of the annotated entity. Consider, for example, the following sentence:

After their debut The Intelligence got signed to *In the Red Records*.

Here, we may infer that the entity *In the Red Records* is a Record Label, thanks to its LAST.FM URL: `http://www.last.fm/label/In+the+Red+Records`. This information is extracted from the whole LAST.FM corpus for those entities falling in one of the four *musical categories* previously defined.

### 3.1. Data Enrichment

For the creation of the ELMD dataset, the crowdsourced annotations extracted from LAST.FM biographies are combined with decisions made by ELVIS and its voting framework.

Every entity mention annotated in the LAST.FM corpus is a candidate to be included in ELMD. The challenge is to assign to each entity its correct DBpedia URI. We approach this problem by leveraging (1) The DBpedia URI assigned by ELVIS, (2) The *agreement score* for that prediction, as well as (3) The *type* information derived from the entity's LAST.FM URL. Our intuition is that the higher the *agreement score*, the more likely the prediction is to be correct. Likewise, we also hypothesize that if a linking decision made by ELVIS coincides in *type* with the original LAST.FM annotation, it is more likely to be correct. Since there is no direct mapping between LAST.FM and DBpedia types, we manually set the type equivalences shown in Table 1.

Regarding the *agreement score*, it corresponds to the number of systems that agreed in a decision (see **Score** column in Table 2). Note that an *agreement score* of 1 may be caused either by cases in which only one system detected an entity mention, or when there is disagreement among systems, but one and only one of them coincides in *type* with the original LAST.FM annotation (last row in Table 2).

As for *type value*, this is a binary value (*type-equivalent* or *type-discrepant*) based on coinciding types between LAST.FM URLs and ELVIS decisions.

## 4. Evaluation

Considering the different possibilities of agreement across the systems integrating ELVIS, there are in total 7 possible configurations: 1 with **full agreement** (score$=3$); 3 with **partial agreement** (score $=2$); and 3 **singleton** configurations (score$=1$). Moreover, considering also the two possible values of *type agreement*, namely `equivalent` and `discrepant`, we have a total number of 14 configurations. Figure 2 provides a visual overview of these con-

| Context | Last.fm type | Tagme | Babelfy | Spotlight | Score | Type Eq. |
|---|---|---|---|---|---|---|
| and the academic minimalism of **Steve Reich** | Artist | Steve_Reich (type:artist) | Steve_Reich (type:artist) | Steve_Reich (type:artist) | 3 | type-equivalent |
| The new album **Hypocrisy** followed shortly thereafter | Album | — | Hypocrisy (type:band) | Hypocrisy (type:band) | 2 | type-discrepant |
| The third album **Lucifer Songs**, opened new and unexpected doors | Album | — | Lucifer_Songs (type:album) | — | 1 | type-equivalent |
| The band's debut album, **Cookies**, was released on 14 May 2007 | Album | HTTP_cookie (type:unknown) | Cookies (type:album) | — | 1 | type-equivalent (only Babelfy) |

Table 2: Agreement examples

| Last.fm type | DBpedia type |
|---|---|
| Song | DBpedia:Song, DBpedia:Single, Yago:Song |
| Album | DBpedia:Album, Yago:Album, Schema:MusicAlbum |
| Artist | DBpedia:MusicalArtist, DBpedia:Band, Schema:MusicGroup, Yago:Musician, Yago:Creator, DBpedia:Artist |
| Record Label | DBpedia:RecordLabel |

Table 1: Type equivalence

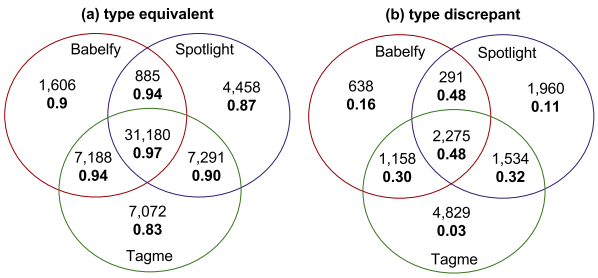


Figure 2: Number of entities and precision of the manual evaluation. Note the major differences in Precision between *type-equivalent* and *type-discrepant* systems.

| | Agreement | Precision | No. Entities |
|---|---|---|---|
| type-equivalent | $= 3$ | 0.97 | 31,180 |
| | $\geq 2$ | 0.96 | 46,544 |
| | $\geq 1$ | 0.94 | 59,680 |
| all | $= 3$ | 0.94 | 33,455 |
| | $\geq 2$ | 0.90 | 51,802 |
| | $\geq 1$ | 0.81 | 72.365 |

Table 3: Precision and number of entities with this value of precision. *Type-equivalent* implies entities from the type-equivalent configuration only, whilst *All* implies all entities regardless their type information.

figurations, where we show both Precision scores for each configuration (in bold) in addition to the number of entities disambiguated with ELVIS in each case.

We evaluated 100 randomly selected entity samples (25 for each of the four Music categories we consider) from each one of the 14 possible configurations, and asked an evaluator with computational linguistics background to manually assess the correctness of the 1,400 predictions. From scores obtained from manual evaluation, we estimated Precision for the whole ELMD dataset with different ranges of *agreement score* as well as two options *type*-wise (see Table 3). The precision value for all the entities is computed proportionally according to the number of entities and the precision obtained in the manual evaluation for the *type-equivalent* and *type-discrepant* settings, hence these can be seen as Micro Average Precision numbers.

We observe that the *type-equivalent* configuration yields much better Precision with only a slight tradeoff in terms of Recall. Therefore, we decided to select for the final ELMD dataset only those URIs stemming from a *type-equivalent* setting where *agreement score* is equal or greater to 1. This ensures a Precision of at least 0,94 in terms of Entity Linking. Moreover, a manual survey of false positives in the highest scoring setting (*agreement score*$= 3$ and *type-equivalent*) showed that these are cases in which even a human annotator may not find it trivial to correctly find the correct entity to those entity mentions. One of these cases are those in which ELVIS is presented with and entity mention that on surface may refer to either an Artist or an Album named after the artist or band itself. An actual case of false positive in our evaluation dataset is the following sentence:

> Her debut album , *Kim Wilde*, (released on RAK records) came out in July 1981 and stayed in the U.K. album charts for 14 weeks, peaking at number 3 and getting much acclaim.

Here, the entity *Kim Wilde* should be disambiguated as the Album with the same name as the artist, but ELVIS incorrectly assigned the Artist's DPpedia URI: `dbpedia.org/resource/Kim_Wilde`. In ELMD there

| Musical Category | Annotations | Distinct Entities | Avg. words | Most frequent entity |
|---|---|---|---|---|
| Song | 3,302 | 2,823 | 2.81 | Shine (6) |
| Album | 7,872 | 6,897 | 2.69 | Like Drawing Blood (6) |
| Artist | 46,337 | 17,535 | 1.88 | The Beatles (160) |
| Record Label | 2,169 | 815 | 1.94 | Sub Pop (33) |

Table 4: Statistics of the linked entities in ELMD. We report, for each *musical category*, the total number of annotations linked to DBpedia, number of unique entities, average number of words per entity mention, and most frequently annotated entity (along with its frequency).

are 50 cases where the same surface text is correctly linked to an Artist entity in some sentences, and to a Song entity in others. Similar ambiguous cases involving Artist and Album (148) and Song and Album (95) are correctly resolved by our system. These particularly challenging cases may be interesting for training Music specific EL algorithms.
Another interesting source of false positives comes between musical entities and equally named entities (not necessarily related to Music). In cases in which the latter are more popular in a reference KB, e.g. their associated node in the graph may have higher connectivity, may become prioritized by disambiguation EL algorithms that consider graph connectivity as a feature. Consider the following sentence:

> He is becoming more and more in demand for his remixing skills; working for the likes of Justin Timberlake and Armand van Helden, and labels including *Ministry Of Sound*, Defected and Intec, to name a just a few.

Here, the entity *Ministry of Sound* refers to a Record Label, a spin-off of the well-known club, which is the entity that was incorrectly assigned: `dbpedia.org/resource/Ministry_of_Sound`. Cases like this would require, first, to ensure that the different entities derived from *Ministry of Sound* (such as the Record Label or a clothing brand of the same name) exist in a reference KB, and second, to exploit contextual information so that a correct decision is made. A similar situation happens when song or album names may be confused with very common words or expressions (e.g. 'Easy', 'Stupid', 'Sad song', 'If', 'Be there'). ELMD is rich in challenging cases like these.

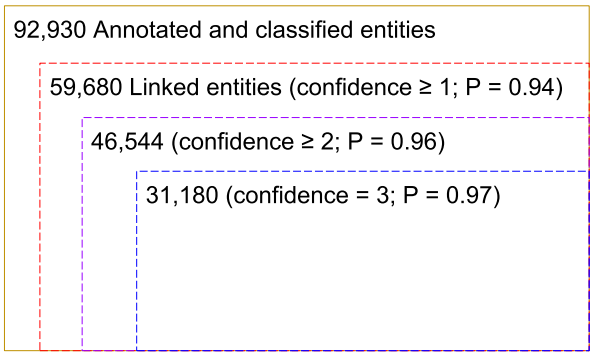


Figure 3: ELMD Overview. Number of entities, confidence score and precision values in different subsets of the dataset.

## 5. Conclusion and Discussion

In this paper we have described two main contributions. First, for the task of Entity Linking, we have presented an integration framework called ELVIS which, based on a voting procedure which leverages decisions made by an arbitrary number of off-the-shelf EL systems, provides high confident entity disambiguations. Currently, ELVIS incorporates three state-of-the-art systems, namely DBpedia Spotlight, Tagme and Babelfy, and can be easily extended with additional systems. The *ELVIS* code is available at https://github.com/sergiooramas/elvis. Second, we have leveraged the potential of ELVIS for the creation of a novel benchmarking dataset for EL in the Music domain, called ELMD. This corpus comes from a collection of LAST.FM artist biographies, and contains 47,254 sentences with 92,930 annotated and classified entity mentions (64,873 Artists, 16,302 Albums, 8,275 Songs and 3,480 Record Labels). From this set of entity mentions, 59,680 are linked to DBpedia (see Table 4), with a precision of at least 0,94. In addition, by setting up a higher confidence threshold it is possible to obtain a subset of ELMD that prioritize higher Precision by sacrificing Recall (see Figure 3). The ELMD dataset together with the evaluation data can be downloaded from http://mtg.upf.edu/download/datasets/elmd.

## Acknowledgements

This work was partially funded by the Spanish Ministry of Economy and Competitiveness under the Maria de Maeztu Units of Excellence Programme (MDM-2015-0502).

## 6. References

Cornolti, M., Informatica, D., Ferragina, P., Informatica, D., and Ciaramita, M. (2013). A Framework for Benchmarking Entity-Annotation Systems. *Proceedings of the International World Wide Web Conference (WWW) (Practice & Experience Track), ACM (2013)*.

Ferragina, P. and Scaiella, U. (2012). Fast and Accurate Annotation of Short Texts with Wikipedia Pages. *Software, IEEE*, 29(1), June.

Gruhl, D., Nagarajan, M., Pieper, J., Robson, C., and Sheth, A. (2009). Context and Domain Knowledge Enhanced Entity Spotting In Informal Text. In *The Semantic Web-ISWC 2009*, pages 260–276. Springer.

Mendes, P. N., Jakob, M., García-Silva, A., and Bizer, C. (2011). DBpedia Spotlight: Shedding Light on the Web of Documents. In *Proceedings of the 7th International Conference on Semantic Systems*, pages 1–8. ACM.

Moro, A., Raganato, A., and Navigli, R. (2014). Entity Linking meets Word Sense Disambiguation: A Unified Approach. *Transactions of the Association for Computational Linguistics*, 2:231–244.

Navigli, R. and Ponzetto, S. P. (2012). BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. *Artificial Intelligence*, 193:217–250.

Oramas, S., Sordo, M., Anke, L. E., and Serra, X. (2015a). A Semantic-Based Approach for Artist Similarity. In *Proceedings of the 16th International Society for Music Information Retrieval Conference, ISMIR 2015, Málaga, Spain, October 26-30, 2015*, pages 100–106.

Oramas, S., Sordo, M., and Espinosa-anke, L. (2015b). A Rule-Based Approach to Extracting Relations from Music Tidbits. *WWW'15 Conference: 2nd Workshop in Knowledge Extraction from Text KET 2015*.

Ostuni, V. C., Di Noia, T., Di Sciascio, E., Oramas, S., and Serra, X. (2015). A Semantic Hybrid Approach for Sound Recommendation. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, pages 85–86. ACM.

Rizzo, G. and Erp, M. V. (2014). Benchmarking the Extraction and Disambiguation of Named Entities on the Semantic Web. *9th International Conference on Language Resources and Evaluation*.

Sordo, M., Oramas, S., and Espinosa-Anke, L. (2015). Extracting Relations from Unstructured Text Sources for Music Recommendation. In *Natural Language Processing and Information Systems*, pages 369–382. Springer.

Tata, S. and Di Eugenio, B. (2010). Generating Fine-Grained Reviews of Songs from Album Reviews. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1376–1385. Association for Computational Linguistics.

Usbeck, R., Roder, M., Ngonga Ngomo, A.-C., and al. (2015). GERBIL - General Entity Annotator Benchmarking Framework. *24th WWW conference, (2015)*.

Zhang, X., Liu, Z., Qiu, H., and Fu, Y. (2009). A Hybrid Approach for Chinese Named Entity Recognition in Music Domain. *2009 Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing*, pages 677–681, December.