PURPOSE OF INVESTIGATION:

There are numerous online tools for entity extraction. Some of them have specified uses, like tagging blog posts; others simply find salient entities and categorise them; others yet are able to identify events and facts. In this investigation, articles about musicians were given to 5 online entity extractors to evaluate their proficiency in identifying names and locations. Later in this project, it might be useful to consider the relationships between entities that these tools sometimes return.

TOOLS & SOURCES:

Extractors tested:
OpenCalais
Extractiv
Alchemy
DBpedia Spotlight
Zemanta

Notes on extractors: OpenCalais, Extractiv, and Alchemy extract entities and categorise them. OpenCalais and Extractiv have the most comprehensive categorisation, and they are also the only two that identify events and facts (e.g. family relations). Zemanta is a blog-tagger, and Spotlight is an annotation tool for DBpedia articles; both do not return categories.

Sources of articles:
Baker's Biographical Dictionary of Musicians
Oxford Dictionary of Music
Wikipedia

Notes on sources: Baker's and Oxford are written in point-form, where as Wikipedia is written in prose. Musicians are chosen across different time periods.

METHOD:

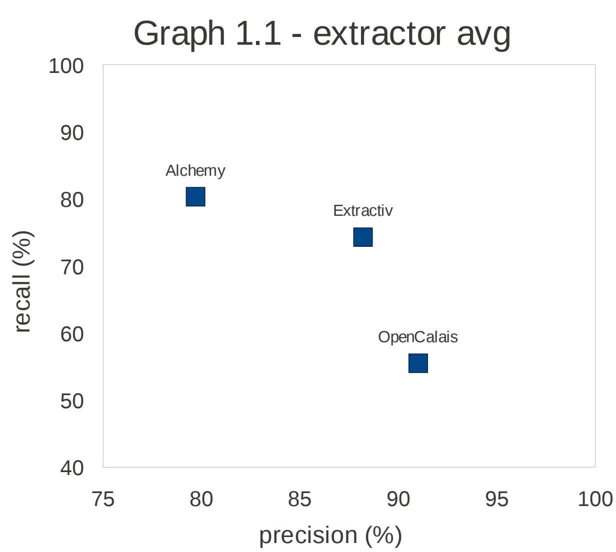
The following ten musicians were used: Anton Bruckner, William Byrd, Arcangelo Corelli, Hans Hassler, Hildegard von Bingen, Gustav Mahler, Palestrina, Ignaz Pleyel, Henry Purcell, and Clara Schumann. Three articles were found for each person (Baker's, Oxford, Wikipedia). Articles from Baker's and Oxford were used in their entirety; articles from Wikipedia were usually cropped to have less than 1000 words. The articles were fed to the extractors and the output recorded. For the extractors that categorised their outputs, the entities returned were recorded under their given categories.

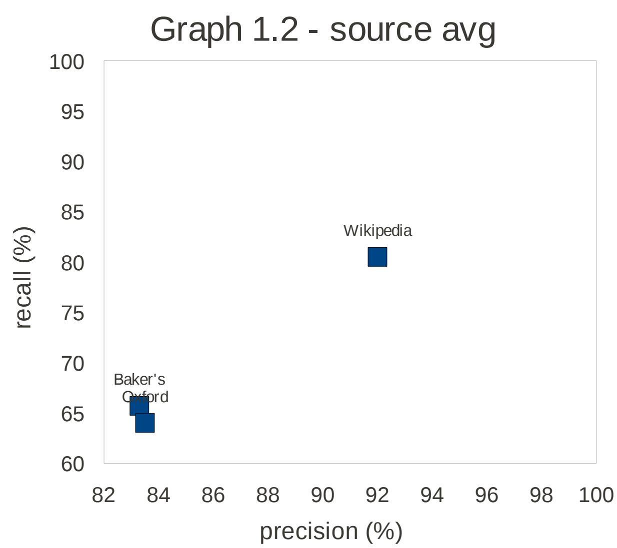
RAW AND PROCESSED DATA

Extracted entities are recorded in individual files for each musician. They can be found in a directory close to this file. A tally of the entities returned can be found in a table in Appendix A. Using the data in these tables, precision and recall values were calculated for each extractor for each source. These values can be found in Appendix B. The averages in the following table are calculated using the values in Appendix B. Sample calculations for all the processed data can be found in Appendix C.

Table 1.7 – Recall and precision averages of totals from different sources:

	Baker's avg		Oxford avg		Wikipedia avg		**extractor avg**	
	recall	precision	recall	precision	recall	precision	**recall**	**precision**
OpenCalais	45.4	90.9	45.8	83.4	75.4	98.6	**55.5**	**91**
Extractiv	67.2	88.2	71.1	86	84.7	90.4	**74.3**	**88.2**
Alchemy	84.4	70.8	75.2	81.1	81.3	87.2	**80.3**	**79.7**
source avg	**65.7**	**83.3**	**64**	**83.5**	**80.5**	**92**		





Notes on graphs: Graph 1.1 uses data from the extractor average columns of table 1.7. Graph 1.2 uses the source average row from the same table.

EVALUATION

DISCUSSION:

As seen in the two graphs above, data from Spotlight and Zemanta were not included. After the data collection stage, it became obvious that these two extractors were not suitable for our purposes. Spotlight is specifically meant to be used as a tool to automatically find links in articles to other DBpedia articles according to Wikipedia linking conventions. Thus, it does not extract entities that are not existing articles on DBpedia (i .e. it wouldn't be able to pick out "Saining Li" from "Saining Li was a famous cartographer of the 16th century"). Also, a large factor affecting whether or not a link is made is how many in-links the target article has. This leads to many extracted entities that are not useful (words like "born"). Despite its shortcomings, it is the most configurable of the extractors. The user is able to regulate precision and recall using the "confidence" field, and also select which categories to include in the extraction process. However, the articles on DBpedia are not very well-organised, and most have unknown categories, making the feature essentially useless at this point. Spotlight has the potential to be a powerful extractor, but it's not there yet. Much less can be said about Zemanta. It's tolerable.

OpenCalais, Extractiv, and Alchemy were used to extract the fields “names”, “locations”, “organisations”, “facilities”, “positions” and (the extremely vague) “other”. The last four fields were difficult to work with, since it's not entirely clear what constitutes a “position”. For example, if “queen” can be categorised as a position, why not “daughter?” Thus, while calculating precision and recall, only locations and names extracted were used. Entity relevance was determined by me, the human. To avoid ambiguity, all names were considered names, even if it was part of a title (e.g. “Tristan” and “Isolde” are both considered names in “Tristan und Isolde”).

ANALYSIS OF GRAPHS:

Graph 1.1 indicates that Extractiv is the best extractor to use. Assuming a linear relationship between recall and precision, a line can be drawn between the points between OpenCalais and Alchemy. Since Extractiv lies above this line, it can be said to have the best performance. This assumes that recall and precision have equal weights. A crude measurement, but simple.

Graph 1.2 shows that recall and precision are both around 10% higher when extraction is performed on Wikipedia than on Oxford and Bakers. It is possible that the extractors used Wikipedia as a training corpus, thus explaining the better performance. However, it is also possible that the extractors simply perform better on prose. Oxford and Baker's are both dictionaries, and as such, employ dictionary abbreviations. As an example, “conservatory” is shortened to “cons.” When the sentence “she was teacher of pf.-playing in the Hoch cons.” (Baker's) is given to OpenCalais, it is unable to identify “Hoch cons.” When “cons.” is changed to “conservatory,” it is able to correctly identify it as a facility. Also, Baker's and Oxford both often refer to people referenced inside their articles solely by their last names, where as full names are usually used on Wikipedia (for the first mention, at least). For example, the sentence “His comp.s, in which Wagner's influence is strongly felt, include 9 symphonies” (Baker's) yields no entities in OpenCalais. However, adding “Richard” before “Wagner” allows OpenCalais to correctly identify “Richard Wagner” as a person. Adding the sentence “Richard Wagner liked fruit.” before the original sentence accomplishes the same thing.

INDIVIDUAL EXTRACTOR FEATURES:

OpenCalais works amazingly well with business-related articles (e.g. it can identify companies, mergers, and acquisitions accurately). It's also one of the only two extractors that identify relationships (Extractiv is the other).

Extractiv is the only extractor that attempts to extract dates. It rarely links these dates with events, and sometimes does not identify years as dates, though.

Alchemy doesn't have too much going for it.

SOURCES OF ERROR:

A probable source of error is human error in identifying all relevant entities. Although I tried to go through the articles very thoroughly, I'm sure I missed some names and locations. As a result, recall values might be higher than they should be.

APPENDIX A - RAW DATA:

Table 1.1 – Raw data of articles from Baker's

		Human	OpenCalais		Extractiv		Alchemy	
			all	relevant	all	relevant	all	rel
Bruckner	names	5	1	1	3	3	5	4
	loc.	5	3	3	5	4	3	3
	total	10	4	4	8	7	8	7
Byrd	names	9	3	2	3	3	11	9
	loc.	2	0	0	2	2	2	2
	total	11	3	2	5	5	13	11
Corelli	names	13	2	2	10	10	18	13
	loc.	10	6	6	7	7	10	10
	total	23	8	8	17	17	28	23
Hassler	names	13	6	6	6	5	26	13
	loc.	7	5	4	4	4	4	4
	total	20	11	10	10	9	30	17
Hildegard	names							
	loc.							
	total							
Mahler	names	8	3	3	4	4	13	7
	loc.	7	3	3	6	6	3	3
	total	15	6	6	10	10	16	10
Palestrina	names	18	13	11	19	12	25	17
	loc.	5	3	2	5	4	6	5
	total	23	16	13	24	16	31	22
Pleyel	names	3	1	1	3	2	3	2
	loc.	6	4	4	4	4	4	4
	total	9	5	5	7	6	7	6
Purcell	names	16	10	7	17	15	28	16
	loc.	7	5	5	7	4	5	5
	total	23	15	12	24	19	33	21
Schumann	names	7	3	3	6	5	21	7
	loc.	6	5	5	7	6	4	4
	total	13	8	8	13	11	25	11

Table 1.2 – Raw data of articles from Oxford

		Human	OpenCalais		Extractiv		Alchemy	
			all	relevant	all	relevant	all	rel
Bruckner	names	12	3	3	9	9	14	10
	loc.	7	5	5	7	7	6	6
	total	19	8	8	16	16	20	16
Byrd	names	7	3	0	8	7	8	6
	loc.	4	1	1	3	3	3	3
	total	11	4	1	11	10	11	9
Corelli	names	6	0	0	5	4	6	6
	loc.	6	2	2	2	2	3	3
	total	12	2	2	7	6	9	9
Hassler	names	2	1	1	2	2	2	2
	loc.	5	4	4	2	2	2	2
	total	7	5	5	4	4	4	4
Hildegard	names	25	7	7	33	25	29	20
	loc.	11	5	5	9	7	8	7
	total	36	12	12	42	32	37	27
Mahler	names	19	4	4	13	10	19	14
	loc.	13	10	9	15	13	10	10
	total	32	14	13	28	23	29	24
Palestrina	names	5	5	4	5	4	9	5
	loc.	3	1	1	3	3	2	2
	total	8	6	5	8	7	11	7
Pleyel	names	6	5	5	4	4	9	6
	loc.	5	4	4	6	5	5	5
	total	11	9	9	10	9	14	11
Purcell	names	17	10	9	14	11	12	9
	loc.	3	1	1	2	2	2	1
	total	20	11	10	16	13	14	10
Schumann	names	6	2	1	4	2	7	5
	loc.	6	5	5	2	2	3	3
	total	12	7	6	6	4	10	8

Table 1.3 – Raw data of articles from Wikipedia

		Human	OpenCalais		Extractiv		Alchemy	
			all	relevant	all	relevant	all	rel
Bruckner	names	6	6	6	6	6	6	6
	loc.	1	0	0	1	1	1	1
	total	7	6	6	7	7	7	7
Byrd	names	8	5	5	9	8	10	7
	loc.	2	2	2	2	2	2	2
	total	10	7	7	11	10	12	9
Corelli	names	13	8	8	9	9	10	10
	loc.	10	9	9	8	8	10	10
	total	23	17	17	17	17	20	20
Hassler	names	12	9	9	8	8	12	10
	loc.	9	6	6	10	9	6	6
	total	21	15	15	18	17	18	16
Hildegard	names	8	6	6	12	6	10	8
	loc.	5	3	3	3	3	2	2
	total	13	9	9	15	9	12	10
Mahler	names	8	6	6	8	8	8	6
	loc.	6	6	6	8	5	5	5
	total	14	12	12	16	13	20	11
Palestrina	names	9	9	8	10	9	10	9
	loc.	8	5	5	10	8	5	5
	total	17	14	13	20	17	15	14
Pleyel	names	8	6	6	6	6	7	6
	loc.	3	3	3	3	3	4	3
	total	11	9	9	9	9	11	9
Purcell	names	13	12	11	10	9	14	13
	loc.	4	3	3	3	3	0	0
	total	17	15	14	13	12	14	13
Schumann	names	27	15	15	20	20	24	17
	loc.	6	4	4	6	6	4	4
	total	33	19	19	26	26	28	21

APPENDIX B - PROCESSED DATA:

Table 1.4 – precision and recall using entities extracted from Bakers

		OpenCalais		Extractiv		Alchemy	
		Recall (%)	precision (%)	recall (%)	precision (%)	recall (%)	precision (%)
Bruckner	names	20	100	60	100	80	80
	location	6	100	80	80	60	100
	total	**40**	**100**	**70**	**87.5**	**70**	**87.5**
Byrd	names	22.2	66.7	33.3	100	100	81.8
	location	0	na	100	100	100	100
	total	**18.2**	**66.7**	**45.5**	**100**	**100**	**84.5**
Corelli	names	15.4	100	76.9	100	100	72.2
	location	60	100	70	100	100	100
	total	**34.8**	**100**	**73.9**	**100**	**100**	**82.1**
Hassler	names	46.2	100	38.5	83.3	100	50
	location	57.1	80	57.1	100	57.1	100
	total	**50**	**90.9**	**45**	**90**	**85**	**56.7**
Hildegard	names						
	location						
	total						
Mahler	names	37.5	100	50	100	87.5	53.8
	location	42.9	100	85.7	100	42.9	100
	total	**40**	**100**	**66.7**	**100**	**66.7**	**62.5**
Palestrina	names	61.1	84.6	66.7	63.2	94.4	68
	location	40	66.6	80	80	100	83.3
	total	**56.5**	**81.2**	**69.7**	**66.7**	**95.7**	**71**
Pleyel	names	33.3	100	66.7	66.7	66.7	66.7
	location	66.7	100	66.7	100	66.7	100
	total	**55.6**	**100**	**66.7**	**85.7**	**66.7**	**85.7**
Purcell	names	43.8	70	93.8	88.2	100	57.1
	location	71.4	100	57.1	57.1	71.4	100
	total	**52.2**	**80**	**82.6**	**79.2**	**91.3**	**63.6**
Schumann	names	42.9	100	71.4	83.3	100	33.3
	location	83.3	100	100	85.7	66.7	100
	total	**61.5**	**100**	**84.6**	**84.6**	**84.6**	**44**

Table 1.5 – precision and recall using entities extracted from Oxford

		OpenCalais		Extractiv		Alchemy	
		recall (%)	precision (%)	recall (%)	precision (%)	recall (%)	precision (%)
Bruckner	names	25	100	75	100	83.3	71.4
	location	71.4	100	100	100	85.7	100
	total	**42.1**	**100**	**84.2**	**100**	**84.2**	**80**
Byrd	names	0	0	100	87.5	85.7	75
	location	25	100	75	100	75	100
	total	**9.09**	**25**	**90.9**	**90.9**	**81.8**	**81.8**
Corelli	names	0	na	66.7	8	100	100
	location	33.3	100	33.3	100	50	100
	total	**16.7**	**100**	**50**	**85.7**	**75**	**100**
Hassler	names	50	100	100	100	100	100
	location	80	40	40	100	40	100
	total	**71.4**	**57.1**	**57.1**	**100**	**57.1**	**100**
Hildegard	names	28	100	100	75.7	80	69
	location	45.5	100	63.6	77.8	63.6	87.5
	total	**33.3**	**100**	**88.9**	**76.2**	**75**	**73**
Mahler	names	21.1	100	52.6	67.9	73.7	73.7
	location	69.2	90	100	86.6	77	100
	total	**40.6**	**92.9**	**71.9**	**82.1**	**75**	**82.8**
Palestrina	names	80	80	80	80	100	55.6
	location	33.3	100	100	100	66.7	100
	total	**62.5**	**83.3**	**87.5**	**87.5**	**87.5**	**63.6**
Pleyel	names	83.3	100	66.6	100	100	66.7
	location	80	100	100	83.3	100	100
	total	**81.8**	**100**	**81.8**	**90**	**100**	**78.6**
Purcell	names	52.9	90	64.7	78.7	52.9	75
	location	33.3	100	66.7	100	33.3	50
	total	**50**	**90**	**65**	**81.2**	**50**	**71.4**
Schumann	names	16.7	50	33.3	50	83.3	71.4
	location	83.3	100	33.3	100	50	100
	total	**50**	**85.7**	**33.3**	**66.6**	**66.7**	**80**

Table 1.6 – precision and recall using entities extracted from Wikipedia

		OpenCalais		Extractiv		Alchemy	
		recall (%)	precision (%)	recall (%)	precision (%)	recall (%)	precision (%)
Bruckner	names	100	100	100	100	100	100
	location	0	na	100	100	100	100
	total	**85.7**	**100**	**100**	**100**	**100**	**100**
Byrd	names	62.5	100	100	88.8	87.5	70
	location	100	100	100	100	100	100
	total	**70**	**100**	**100**	**90.9**	**90**	**75**
Corelli	names	61.5	100	69.2	100	76.9	100
	location	90	100	80	100	100	100
	total	**73.9**	**100**	**73.9**	**100**	**86.9**	**100**
Hassler	names	75	100	66.7	100	83.3	83.3
	location	66.7	100	100	90	66.6	100
	total	**71.4**	**100**	**80.6**	**94.4**	**76.2**	**88.8**
Hildegard	names	75	100	75	50	100	80
	location	60	100	60	100	40	100
	total	**69.2**	**100**	**69.2**	**60**	**76.9**	**83.3**
Mahler	names	75	100	100	100	75	100
	location	100	100	83.3	62.5	83.3	62.5
	total	**85.7**	**100**	**92.9**	**81.2**	**78.6**	**81.3**
Palestrina	names	88.9	88.9	100	90	100	90
	location	62.5	100	100	80	62.5	100
	total	**76.5**	**92.9**	**100**	**85**	**82.3**	**93.3**
Pleyel	names	75	100	75	100	75	85.8
	location	100	100	100	100	100	75
	total	**81.2**	**100**	**81.2**	**100**	**81.8**	**81.9**
Purcell	names	84.6	91.7	69.2	90	100	92.9
	location	75	100	75	100	0	NA
	total	**82.4**	**93.3**	**70.6**	**92.3**	**76.5**	**92.9**
Schumann	names	55.6	100	74.1	100	63	70.8
	location	66.7	100	100	100	66.7	100
	total	**57.6**	**100**	**78.8**	**100**	**63.6**	**75**

APPENIX C - SAMPLE CALCULATIONS:

recall = relevant entities extracted / all relevant entities * 100%
= relevant OpenCalais entity / human extracted entity * 100%
= 1/5 * 100 %
= 20.0%

precision = relevant entities extracted / all entities extracted * 100%
= relevant OpenCalais entity / all OpenCalais entities * 100%
= 1/1*100%
= 100%

Baker's recall avg = (Baker's OpenCalais avg recall + Baker's Extractiv avg recall + Baker's Alchemy avg recall) / 3
= (45.4% + 67.2% + 84.4%) / 3
= 65.7%

OpenCalais recall avg = (OpenCalais Baker's avg recall + OpenCalais Oxford avg recall + OpenCalais Wikipedia agv recall) / 3
= (45.4% + 45.8% 75.4%) / 3
= 55.5%

* All calculations attempt to preserve three significant digits