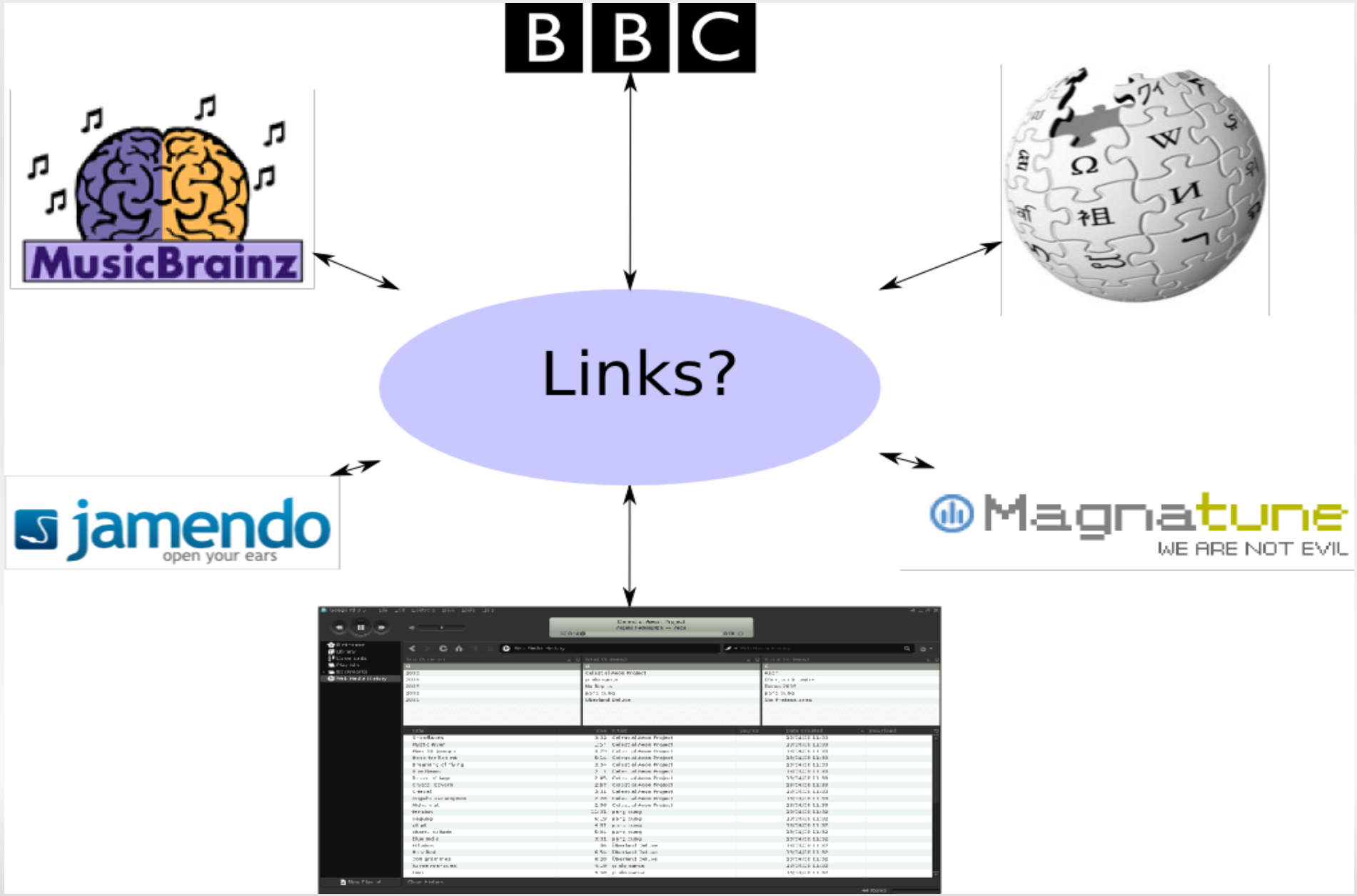
Automatic Interlinking of music datasets on the Semantic Web

Yves Raimond, Christopher Sutton, Mark Sandler
Centre for Digital Music
Queen Mary, University of London
LDOW 2008, 22th of April

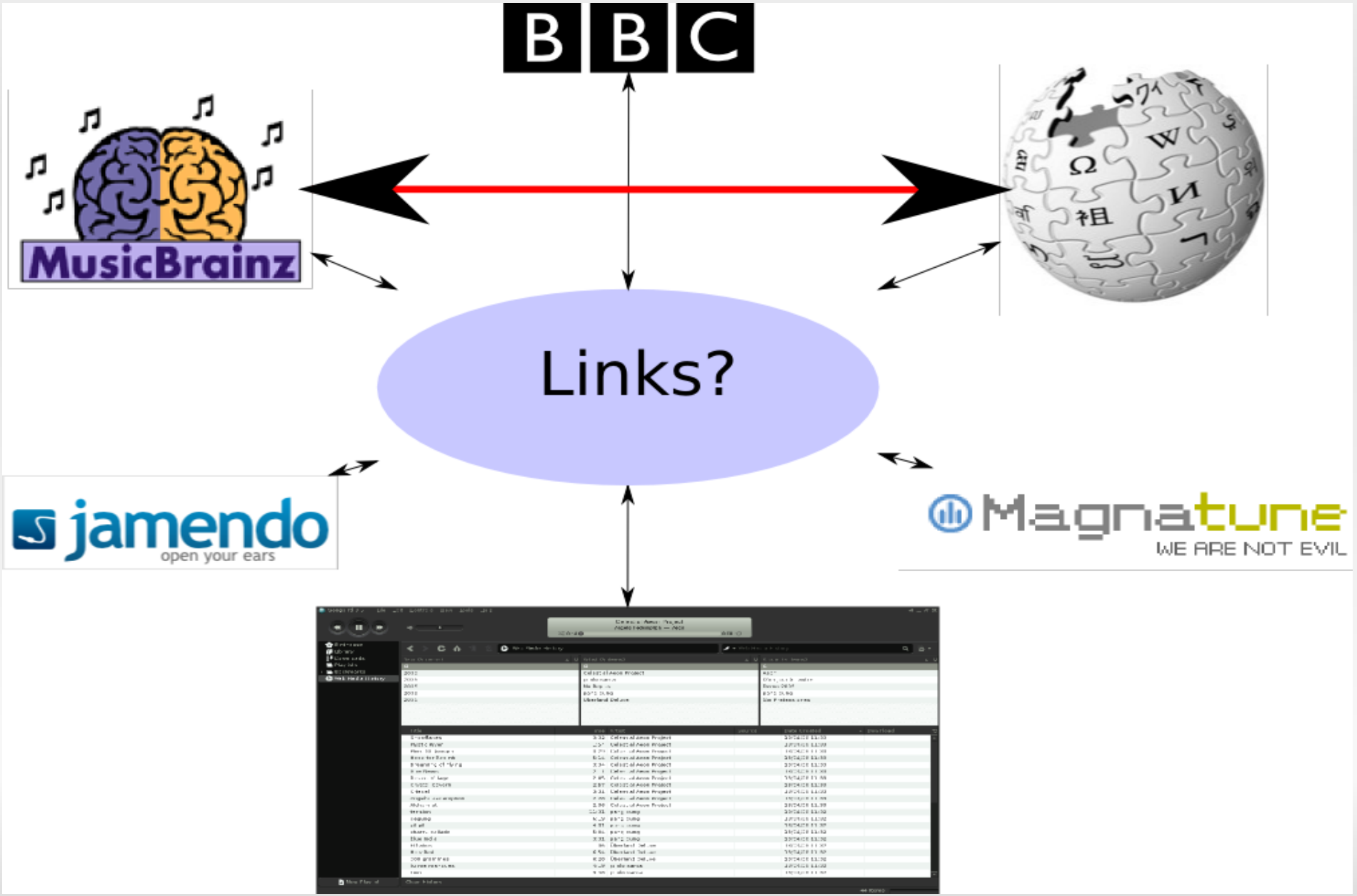
Linked Data publishing

- D2R, Virtuoso
- P2R
- Triplify
- Pubby or URISpace + SPARQL end-point
- API wrappers:
  - RDF Book Mashup
  - Last.fm or MySpace on DBTune
  - Virtuoso Sponger
- Vim and .htaccess :-)

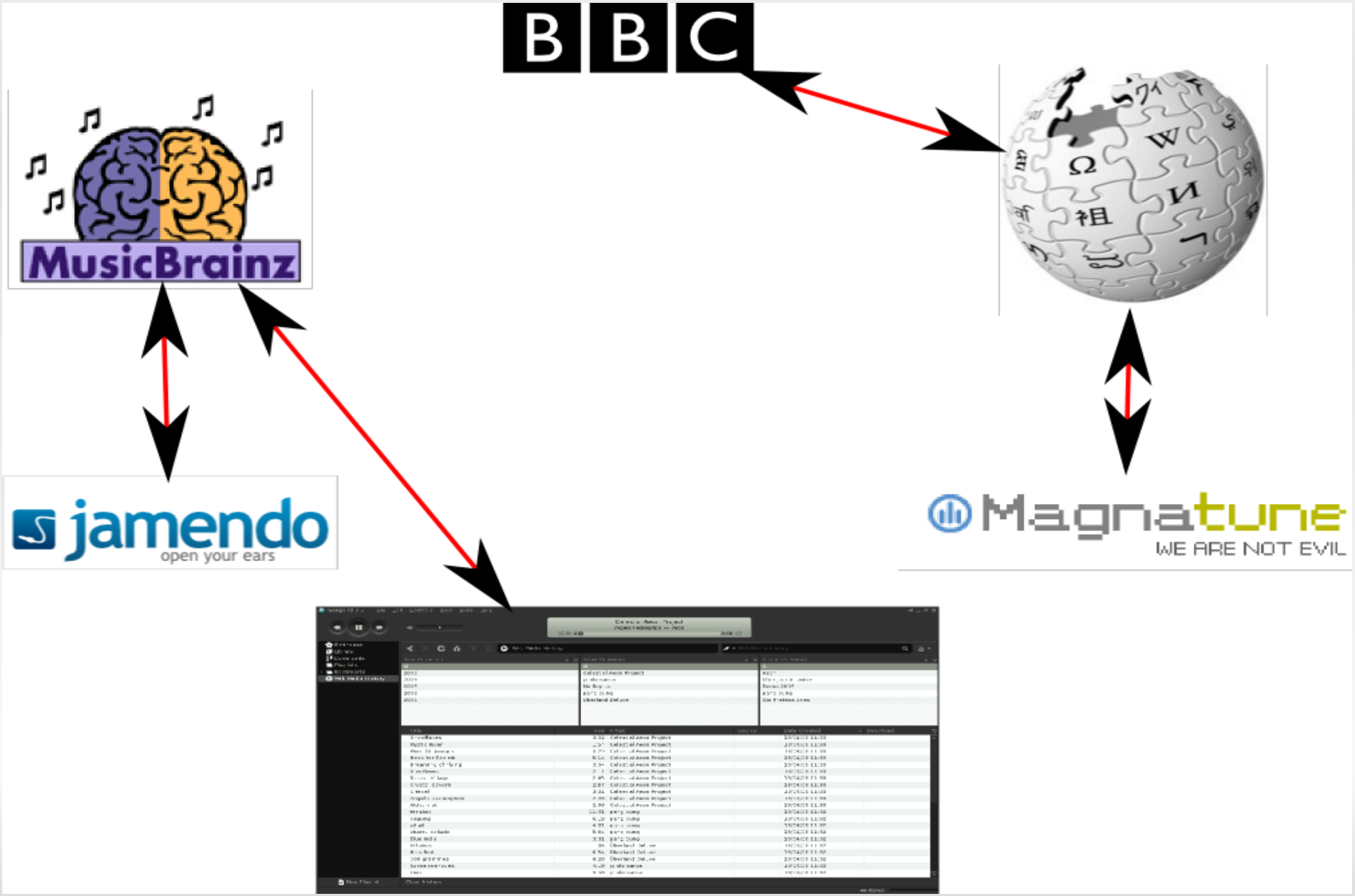
And now?


BBC
MusicBrainz
Links?
jamendo
open your ears
Magnatune
WE ARE NOT EVIL

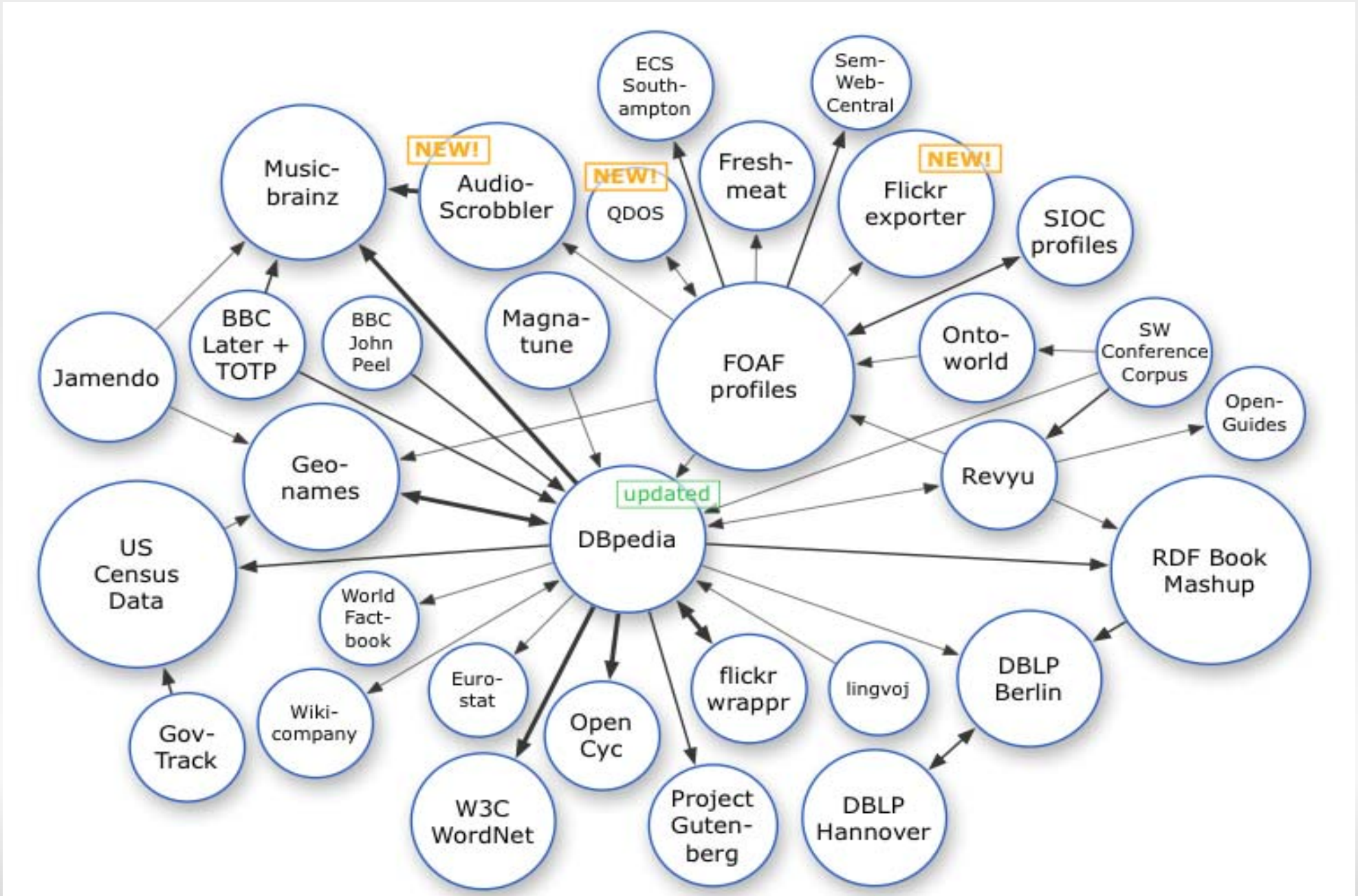
Communities can be helpful



Algorithms can be helpful too



In context

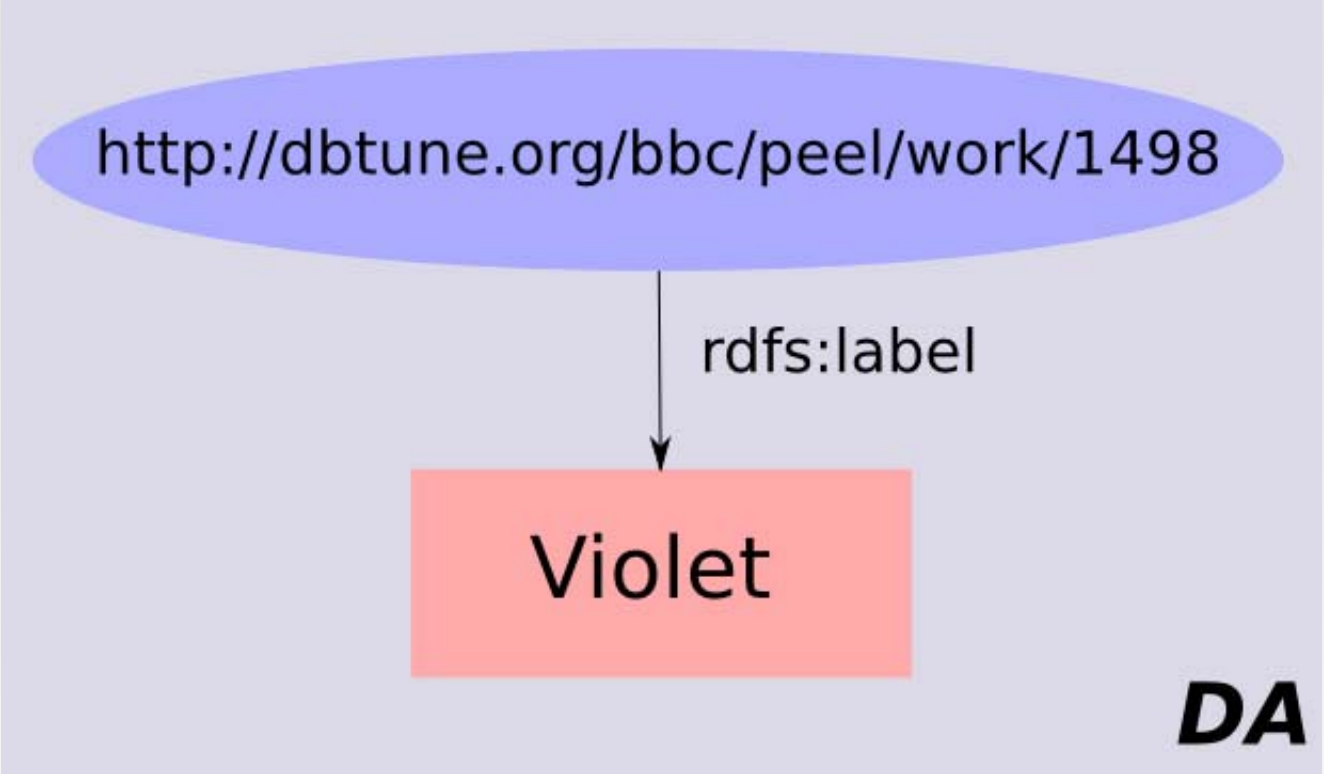

ECS South-ampton
Sem-Web-Central
Music-brainz
NEW!
Audio-Scrobbler
NEW!
QDOS
Fresh-meat
NEW!
Flickr exporter
SIOC profiles
BBC Later + TOTP
BBC John Peel
Magna-tune
FOAF profiles
Onto-world
SW Conference Corpus
Jamendo
Open-Guides
Geo-names
updated
DBpedia
Revyu
US Census Data
RDF Book Mashup
World Fact-book
Euro-stat
flickr wrappr
lingvoj
DBLP Berlin
Wiki-company
Gov-Track
Open Cyc
W3C WordNet
Project Guten-berg
DBLP Hannover

Problem

- Automatically find the overlapping parts between two datasets ***DA*** and ***DB***
  - http://zitgist.com/music/artist/0781a3f3-645c-45d1-a84f-76b4e4dec and http://dbtune.org/jamendo/artist/5
  - http://zitgist.com/music/record/fade0242-e1f0-457b-99de-d9fe0c8c and http://dbtune.org/jamendo/record/33
- Publish corresponding owl:sameAs links
- We want a really low rate of false-positives
  - Violet performed by Hole in a John Peel session IS NOT the same as the flower
  - The French band Both is not the same as the American one

Automatic interlinking – Try 1

- Simple literal lookups



- Query ***DB*** using such labels

Automatic interlinking – Try 1

Violet may refer to:

Color

- Violet (color), which is primarily used to describe a color often confused with purple.

People

- Violet (name), a given name for girls
- Violet Berlin, the presenter of the television program *Game Pad*
- Violet Bonham Carter, British politician
- Violet Wilkey, American child actress of the silent film era
- **Violet**, the stage name of singer Amelia Brightman

In biology

- Violet, a plant of the genus *Viola*
- African violet, a plant of the genus *Saintpaulia* that has a superficial resemblance to *Viola*
- Dogtooth violet, any of several species of the genus *Erythronium*
- Sea violet, a type of edible ascidian, or sea squirt

In geography

- Violet, Louisiana, a city in St. Bernard Parish, Louisiana

In music

- *Violet* (musical), an off-Broadway musical
- *Violet* (album), an album by The Birthday Massacre
- "Violet" (song), a song by the band Hole

Automatic interlinking – Try 2

- Let's restrict the range of the resources we're looking for...

```
PREFIX p: <http://dbpedia.org/property/>
SELECT ?r WHERE {
?r ?p "Violet"@en.
?r a <http://dbpedia.org/class/yago/Song107048000> }
```

Description of http://dbpedia.org/resource/Violet_%28song%29:

property	hasValue	
rdf:type	dbpedia:class/yago/Song107048000	-
rdfs:comment	""Violet" is a song by the rock band Hole, fronted by Courtney Love. It was the third single to be released from their second album, Live Through This, following Doll Parts which was released in 1994. This song was released after an extensive touring period by the band, throughout 1994 and 1995.""@en	-
rdfs:comment	"Violet (fiołek)- jest to trzeci singel z drugiego albumu grunowego zespołu Hole. Został on napisany przez Courtney Love.""@pl	-
skos:subject	:Category:Hole_songs	-
dbpedia2:abstract	""Violet" is a song by the rock band Hole, fronted by Courtney Love. It was the third single to be released from their second album, Live Through This, following Doll Parts which was released in 1994. This song was released after an extensive touring period by the band, throughout 1994 and 1995. "Violet" is a song reputedly about Billy Corgan with whom Courtney had a relationship in the early nineties. Courtney introduced the song on the Jools Holland show in 1995 as "This song's about a jerk, I hexed him and now he's losing his hair," which might refer to Corgan's hair loss. Lyrics such as 'When they get, what they want, then they never want it again', and 'Go on, take everything, take everything, I want you to', portray the end of a bitter relationship. In any event, Corgan and Love subsequently came to enjoy cordial relations. The songwriting credits on record goes collectively to Hole, though BMI's website shows that "Violet" was written only by Love and bandmate Eric Erlandson. "He Hit Me (And It Felt Like A Kiss)" is originally written by Carole King and Gerry Goffin, and was first recorded by producer Phil Spector in the 1960s.""@en	-
dbpedia2:abstract	"Violet (fiołek)- jest to trzeci singel z drugiego albumu grunowego zespołu Hole. Został on napisany przez Courtney Love.""@pl	-
rdfs:label	"Violet (song)"@en	-
rdfs:label	"Violet"@pl	-
dbpedia2:wordnet_type	<http://www.w3.org/2006/03/wn/wn20/instances/synset-phonograph_record-noun-1>	-
dbpedia2:reference	<http://www.foxnews.com/story/0,2933,200533,00.html>	-

Automatic interlinking – Try 2

- Problems:
  - Manually defining constraints is painful
  - They are two artists named ”Both” in Musicbrainz
  - Two songs titled ”Mad Dog” in Dbpedia (by Elastica and Deep Purple)
  - Etc. etc.

Graph matching algorithm

- An algorithm to match a whole RDF graph in ***DA*** to a whole graph in ***DB***
- Intuitive idea:

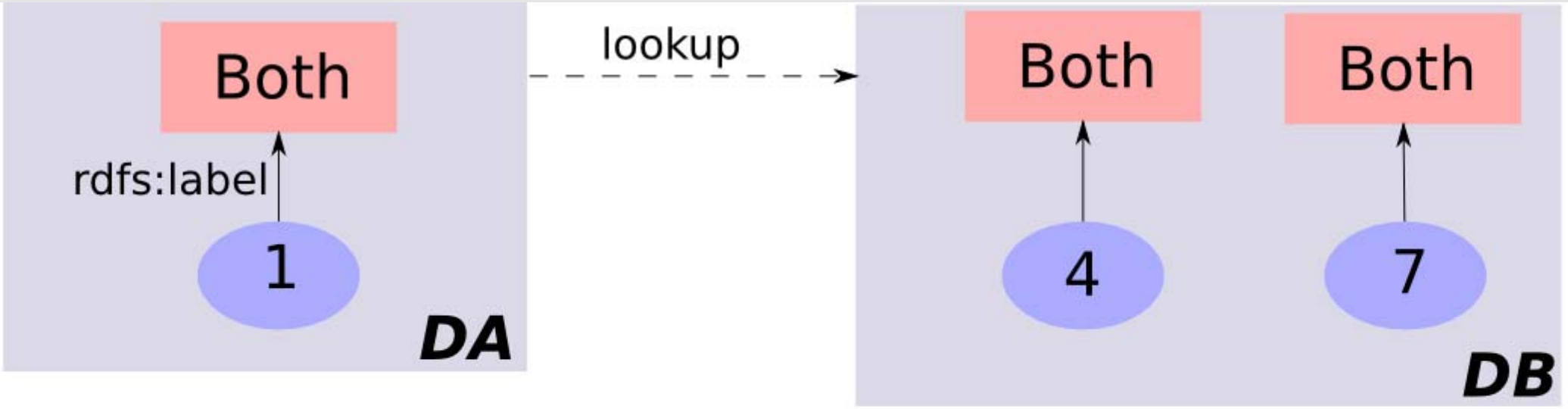
  Two artists that made albums titled similarly are likely to be similar. If the tracks on these albums are titled similarly, they are even more likely to be similar. Etc.
- We explore linked data as long as we don't have enough clues
- Full pseudo-code in the paper

Step 0 – Starting point

- We pick a resource in ***DA***

Step 1 - Lookup

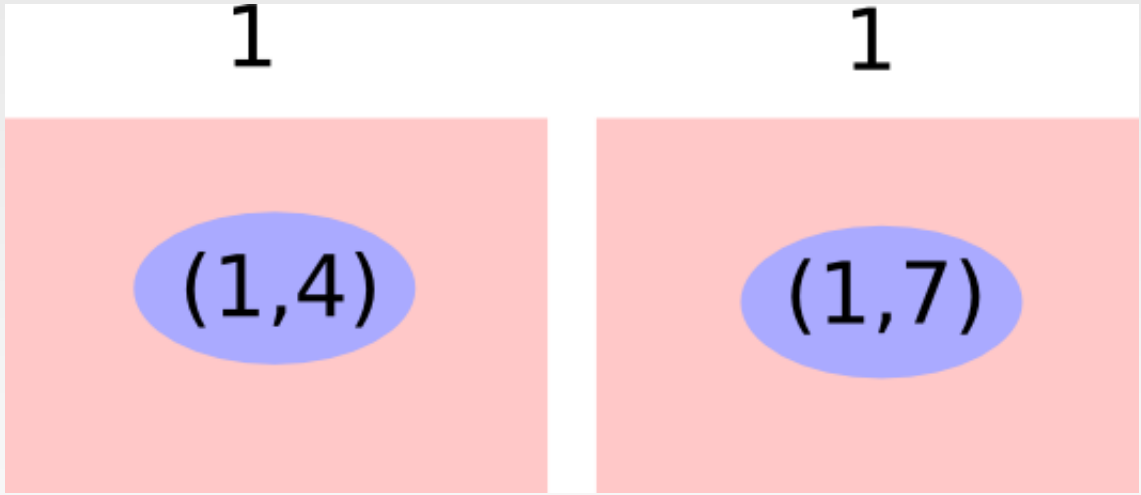
- Dereference starting resource, extract a label
- Lookup ***DB*** as in Try 1 or 2



1: http://dbtune.org/jamendo/artist/5
4: http://zitgist.com/music/artist/5f9f2dfb-76f0-4872-ad7d-f9d84a908cb5
7: http://zitgist.com/music/artist/0781a3f3-645c-45d1-a84f-76b4e4decf6d

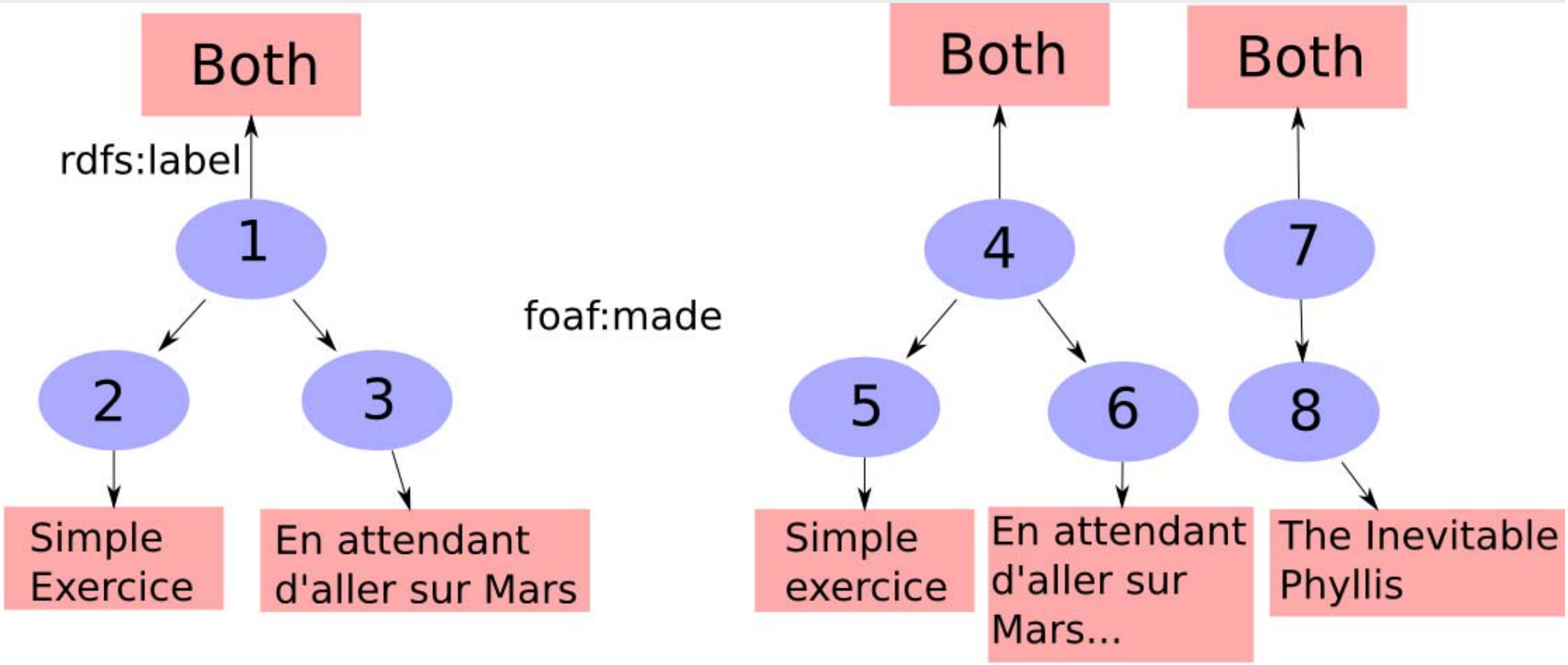
Step 2 – Similarity measure

- Derive possible graph mappings
- Sum of the corresponding resource similarities, normalised by the number of nodes in the graph mapping



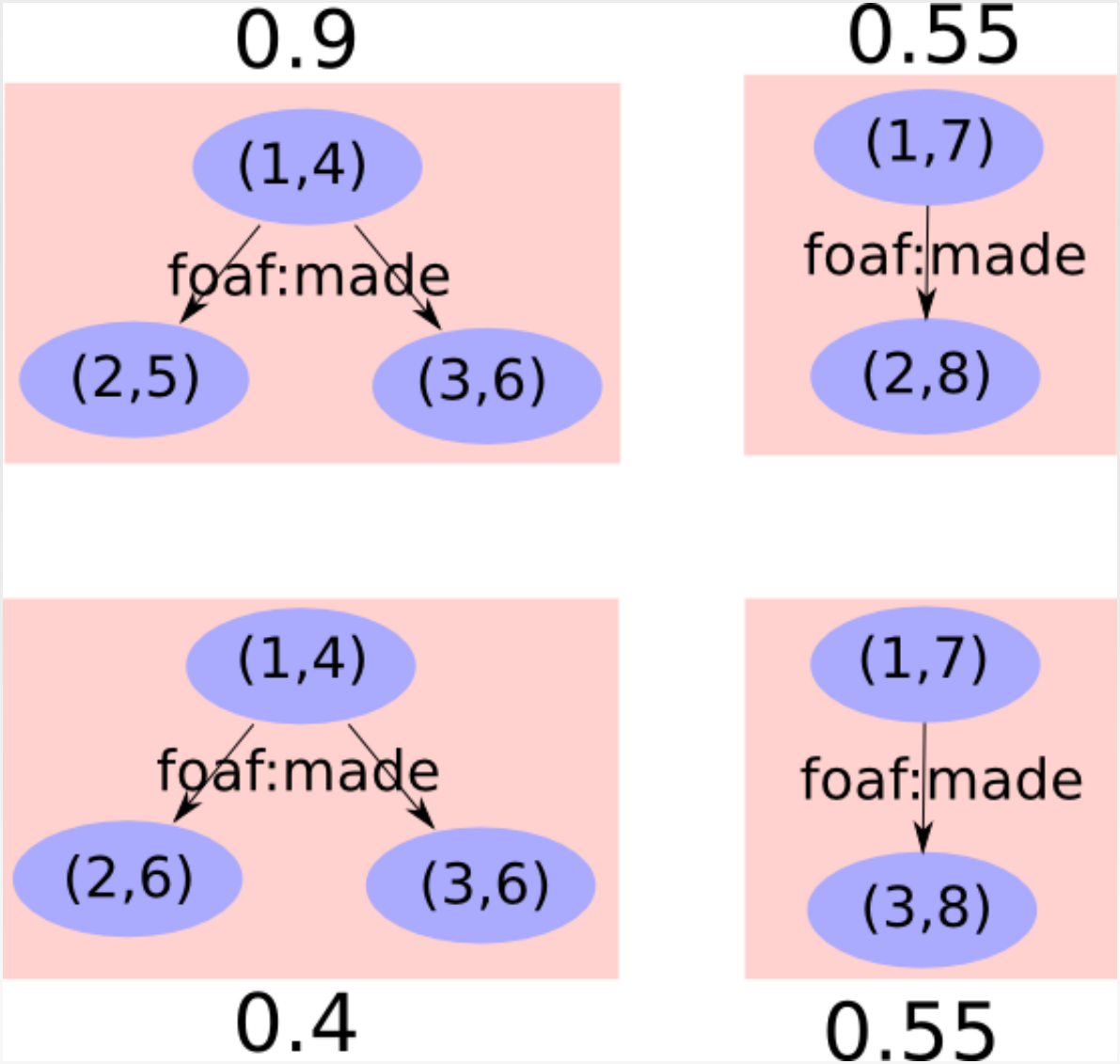
Two above the similarity threshold, we can't make a choice

Step 3 – Explore



2: http://dbtune.org/jamendo/record/33
3: http://dbtune.org/jamendo/record/174
5: http://zitgist.com/music/record/fade0242-e1f0-457b-99de-d9fe0c8cbd57
6: http://zitgist.com/music/record/3042765f-67ba-49ef-ab28-45805fabef4a
8: http://zitgist.com/music/record/2160c817-602c-4ec7-b14b-e7de819e29b6

Step 4 – Update similarity

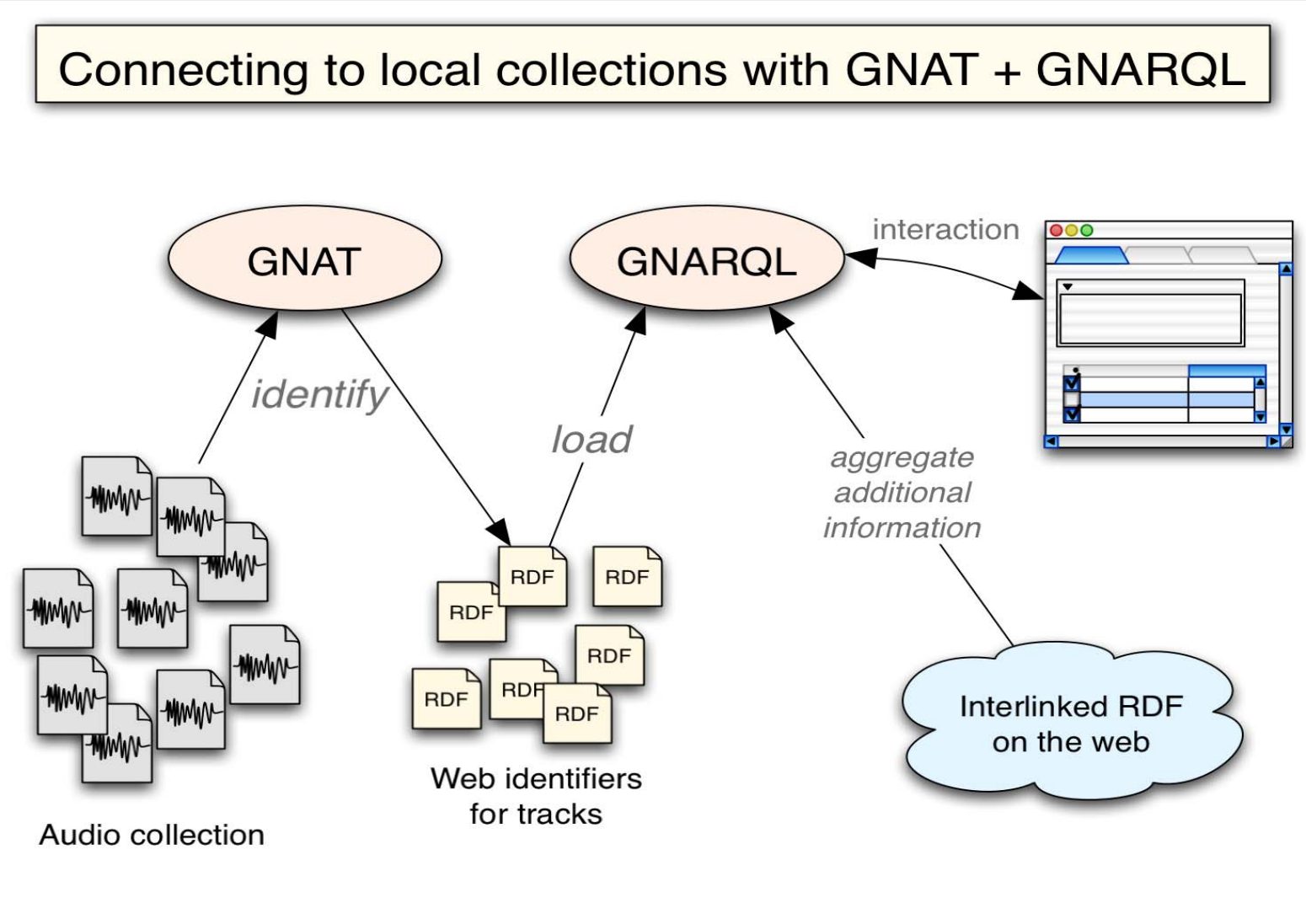


One above our similarity threshold, we make a choice

Experiment 1

- Linking Jamendo to Musicbrainz
  - Prolog implementation (Idmapper in the motools sourceforge project)
  - Evalution: manually checking 60 linkage
    - No incorrect links drawn
    - 53 links not drawn (no matching artists in Musicbrainz)
    - 5 correct links drawn
    - 2 links not drawn that should have been drawn
      - Due to the fact that the RDF version of Musicbrainz is outdated
- Example

Experiment 2



- Evaluation of GNAT in the paper
- Demo

Questions?