vista TV

SEVENTH FRAMEWORK PROGRAMME

ViSTA-TV: Video Stream Analytics for Viewers in the TV Industry

FP7 STREP ICT-296126 | 296126 co-funded by the European Commission
ICT-2011-SME-DCL | SME Initiative on Digital Content and Languages

D6.1 External data service design

Valentina Maccatrozzo (VUA),
Lora Aroyo (VUA),
Guus Schreiber (VUA),
Libby Miller (BBC)

Project start date:	June 1^{st}, 2012	Project duration:	24 months
Document identifier:	ViSTA-TV/2012 D6.1	Version:	v1.0
Date due:	01 December 2012	Status:	FINAL version
Submission date:	22 November 2012	Distribution:	RE

ViSTA-TV Consortium

This document is part of a collaborative research project funded by the FP7 ICT Programme of the Commission of the European Communities, grant number 296126. The following partners are involved in the project:

University of Zurich (UZH) - Coordinator
Dynamic and Distributed Information Systems Group (DDIS)
Binzmʹuhlstrasse 14
8050 Zʹurich, Switzerland
Contact person: Abraham Bernstein
E-mail: bernstein@ifi.uzh.ch

Technische Universitʹat Dortmund (TUDo)
Computer Science VIII: Artificial Intelligence Unit
D-44221 Dortmund, Germany
Contact person: Katharina Morik
E-mail: katharina.morik@cs.uni-dortmund.de

Rapid-I GmbH (RAPID-I)
Stockumer Strasse 475
44227 Dortmund, Germany
Contact person: Ingo Mierswa
E-mail: mierswa@rapid-i.com

Zattoo Europa AG (Zattoo)
Eggbʹuhlstrasse 28
CH-8050 Zʹurich, Switzerland
Contact person: Bea Knecht
E-mail: beaknecht@me.com

Vrije Universiteit Amsterdam (VUA)
Web & Media Group, Department of Computer Science, Faculty of Sciences (FEW)
De Boelelaan 1081a
NL-1081 HV Amsterdam, The Netherlands
Contact person: Guus Schreiber
E-mail: guus.schreiber@vu.nl

The British Broadcasting Corporation (BBC)
56 Wood Lane / Centre House - Broadcasting House
UK-W12 7SB Northampton, United Kingdom
Contact person: Chris Newell
E-mail: Chris.Newell@bbc.co.uk

Executive Overview

This document provides an overview of the external sources identified to provide additional data to the one provided by our broadcasters partners. This includes also additional features extracted from WP2. The deliverable provides also the design of the extraction workflows, which will be performed offline.

Contents

1 Introduction

During these first six months of project, WP6 partners worked mainly on the exploration of possible sources to be used by the external data service. The external data service has the main objective of adding useful information to the EPG data made available by our broadcaster partners, in order to improve the feature selection for the recommendations. The selection was performed considering the wide range of possible topics TV programming covers. This led us to take into consideration general knowledge datasets, such as Wikipedia, as well as particular datasets, such as geographical ones. Considering the main purpose of the live recommendations in the project, we then considered news websites, both news agencies and newspapers, in order to be up-to-date with the latest events of possible interest for the users. This includes also Olympics and, more generally sport websites, in order to give support to the creation of the Olympics dataset. Following this rationale, we started our analysis from the Linked Open Data (LOD) Cloud, and later we took into consideration sources that are not covered in the LOD. This sources are mainly newspapers websites, review websites, sports and Olympics websites, other Electronic Program Guide (EPG) providers. Most of these sources do not provide an API to access them, which means that they should be crawled. Regarding this point, we contacted a no-profit organization, called CommonCrawl[1], which is creating an open repository of web crawl data universally accessible. We can access such amount of data from a facility we have here in the Netherlands, called Sara. Sara is an integrated ICT research infrastructure that provides services in the areas of computing, data storage, visualization, networking, cloud and e-Science. Some of the sources we are interested in are not currently in their crawl, so we provided them a list of the ones we would like to have, namely:

- Films, Series & TV Shows
  - `http://www.imdb.com`
  - `http://www.netflix.com`
  - `http://www.rottentomatoes.com`
- News
  - `http://emm.newsbrief.eu/overview.html`
  - `http://www.guardian.co.uk`
  - `http://www.telegraph.co.uk`
  - `http://www.reuters.com`
- Sports & Olympics
  - `http://www.sportingintelligence.com`
  - `http://www.london2012.com`
- EPGs
  - `www.teleboy.ch/`
  - `www.tele.ch/`
  - `www.tvspielfilm.de/`
  - `www.tvtoday.de/`
  - `http://www.tvprogramm.sf.tv/`
  - `http://www.ard.de/`
  - `http://www.zdf.de/`

Some of these sources are not available to be crawled, however CommonCrawl will try to find an arrangement. These sources will be available in the next crawl which should be performed by the end of the year.

In Section 2 we present an overview of the selected external datasets, in Section 3 we show statistical analysis of those sources. Some examples of enrichments are presented in Section 4, while in Section 5 we describe the design of the data extraction workflows.

[1] `http://commoncrawl.org/`

2 Identification of candidate external datasets and survey their API and rights

2.1 Program metadata

2.1.1 Datasets from Linked Open Data Cloud

Datasets from the Linked Open Data cloud that are the most suitable for the external data service. The selection of these sources was performed considering the wide range of topics a television program can propose.

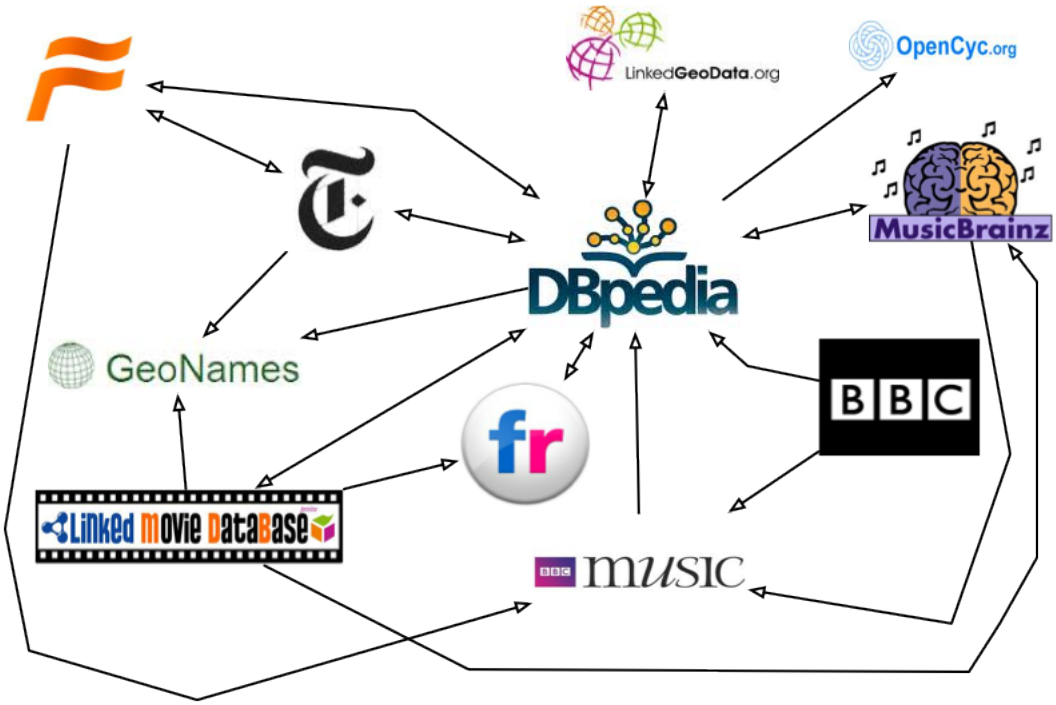

Figure 1: Interlinking of the selected Linked Open datasets

Freebase Freebase is an open database of the world's information. It mainly relies on the work of the community, however there is also a substantial investment from Google. It is free for anyone to query, contribute to, built applications on top of, or integrate into their websites. There are also per-entry RDF pages, and RDF dumps will be available.

License[2]: Creative Commons Attribution

API: Google API Client. Supports many programming languages as Java and Python. Require the use of an API key, which can be easily obtained. The Freebase API Terms of Service limit give users a read quota of 100k API calls per day (rolling 24 hour clock) and a write quota of 10k writes per day. The read services allow to find entities by keyword search, retrieve detailed structured data about entities or collections of entities, get a summary of all the information for an entities, get short textual descriptions for entities, get representative thumbnail images for entities.

Information type:

- standard EPG information (program title, cast, director, genre, etc.)
- awards list
- people relationships (personal and professional ones)
- a large collection of places, sports, entities and general topics suitable as background knowledge

[2]Please find a complete description of the licenses in Appendix A

DBpedia DBpedia is a crowd-sourced community effort to extract structured information from Wikipedia and to make this information available on the Web.

License: Creative Commons Attribution Share-Alike

API: DBpedia lookup service. Two types of search: by keywords or by prefix. It can be run locally as well. The URL has the form `http://lookup.DBpedia.org/api/search.asmx/<API>?<parameters>`. There is also a SPARQL endpoint available at `http://DBpedia.org/sparql`.

Information type:

- standard EPG information
- awards list
- people relationships

LinkedMDB Linked data about movies. Currently the database is not updated.

License: Creative Commons Attribution

API: SPARQL endpoint ia available at http://data.linkedmdb.org/sparql.

Information type:

- standard EPG information
- ratings but not always

NYT For the last 150 years, The New York Times has maintained one of the most authoritative news vocabularies ever developed. In 2009, they began to publish this vocabulary as linked open data.

License: Creative Commons Attribution

API(s): The most relevant ones are:

- The Article Search API: Search Times articles from 1981 to today, retrieving headlines, abstracts and links to associated multimedia.
- The Most Popular API: Get links and metadata for the blog posts and articles that are most frequently e-mailed, shared and viewed by NYTimes.com readers.
- The Movie Reviews API: Get links to reviews and NYT Critics' Picks, and search movie reviews by keyword.
- The Semantic API: get access to the people, places, organizations and descriptors that make up the controlled vocabulary used as metadata by The New York Times.
- The Times Newswire API: Get links and metadata for Times articles in an up-to-the-minute stream.

GeoNames The GeoNames Ontology makes it possible to add geospatial semantic information to the Word Wide Web. All over 6.2 million GeoNames toponyms now have a unique URL with a corresponding RDF web service. Other services describe the relation between toponyms.

License: Creative Commons Attribution

API: No, however it can be downloaded completely.

LinkedGeoData LinkedGeoData is an effort to add a spatial dimension to the Web of Data / Semantic Web. LinkedGeoData uses the information collected by the OpenStreetMap project and makes it available as an RDF knowledge base according to the Linked Data principles. It interlinks this data with other knowledge bases in the Linking Open Data initiative. This dataset has been chosen to be able to perform enrichment of depicted or shot locations.

License: Creative Commons Attribution

API: SPARQL endpoint (available at `http://linkedgeodata.org/`).

MusicBrainz MusicBrainz is an open music encyclopedia that collects music metadata and makes it available to the public.It contains RDF representations of albums, artists, tracks, labels and their relationships. This dataset can be used as a means of communication between DBpedia, BBC, BBC music and IMDB, to link for instance soundtracks and related TV shows. As the previous one, this dataset has been chosen to be able to perform enrichment of depicted or shot locations.

License: Creative Commons Attribution-NonCommercial-ShareAlike

API: SPARQL endpoint (available at `http://dbtune.org/musicbrainz/sparql`).

BBC programmes BBC programmes aims to provide a online guide to the programmes available on BBC channels. They provide updated EPG information about the programmes: description, cast, genre, format and schedule.

License: Creative Commons Attribution

Availability: HTTP request

BBC music BBC music aims to provide a comprehensive guide to music content across the BBC. They are now expanding that service to provide comprehensive information about artists who appear on BBC programmes or who have been covered in bbc.co.uk/music's reviews.

License: Creative Commons Attribution-NonCommercial-ShareAlike

Availability: HTTP request

Dataset	License	Availability
Freebase	Creative Commons Attribution	API
DBpedia	Creative Commons Attribution Share-Alike	SPARQL endpoint
LinkedMDB	Creative Commons Attribution	SPARQL endpoint
NYT	Creative Commons Attribution	API
GeoNames	Creative Commons Attribution	Dump
LinkedGeoData	Creative Commons Attribution	SPARQL endpoint
BBC programmes	Creative Commons Attribution	HTTP request
BBC music	Creative Commons Attribution-NonCommercial-ShareAlike	HTTP request
MusicBrainz	Creative Commons Attribution-NonCommercial-ShareAlike	SPARQL endpoint

Table 1: Summary of the properties of the selected LOD datasets

2.1.2 Project broadcasters

Zattoo Zattoo uses as external sources:

- teletext content: only from DVB-SI signal and are not relying on any other external providers ($e.g.$ websites). Zattoo has a limited set of channels where it provides that service.
- Axel Springer: the German multimedia company provides Zattoo with most of its program data. It imports data for about 90 TV channels on daily basis from their data source. It contains also images and internal and external IDs of programs. Usually the data is of high quality and up to date.
- DVB-SI: for channels it can't get detailed data from Axel Springer it uses DVB-SI (Digital Video Broadcasting - Service Information) data straight from the DVB stream.
- P&T: for its service in Luxembourg Zattoo uses data from their partner P&T Luxembourg since its service there is only available for P&T customers. P&T is a mail and telecommunications company that also offers TV services.
- SF: for the Swiss channels SF1, SF2 and SF info it is directly getting data from them. Since September 2012, it only uses that as a backup for the Axel Springer data.

The data collected is available via web `www.zattoo.com` as well via Zattoo API (zapi). The purposes of the EPG data is to give the user several features:

- tv guide
- search
- current program information on the running stream $e.g.$ `http://zattoo.com/#sf-1`

To obtain the information about channels and EGP data, a HTTPS GET request should be used.

- Channels: GET `https://HOST/zapi/channels`.
- Guide: GET `https://HOST/zapi/program/guide`. Optional parameters are timestamp and channel id.
- Search: GET `https://HOST/zapi/program/search`. Optional parameters are time, query, title, episode title, description and credits.

In the near future, Zattoo will make available a SOLR instance that will be queried for any selection of EPG data.

BBC The objective of using external data within BBC is consuming data to add more context to their content. They take the same TV, radio, news content and shine it through a MusicBrainz prism and get the BBC site `http://www.bbc.co.uk/music` or a DBpedia prism and get `http://www.bbc.co.uk/nature`. This gives them maximum exposure to content from as many angles as possible with potential for links outside the BBC. It is also used to actually storing and managing linked data behind the scenes. The Olympics/sport use case was partly about minimizing journalist effort so that they only needed to 'tag' one thing, for example a country's team, and inferencing tagged related things (for example a team member) so the result was lots of different ways of querying and displaying and aggregating the data for not much effort. The external data sources are:

- Links to Wikipedia, DBpedia, MusicBrainz and to other relevant web pages
- Identifiers from MusicBrainz
- Specific data sources for certain events, e.g. IOC data for the Olympics

The purpose is publishing data to be findable (including via search engines) and to re-use within the BBC and for others to re-use to get links back; and in the HTML case, for people to link to.

Summary of published data

- Schedule data per channel for main radio and TV stations (including regional variants but not local radio) in RDF, JSON, XML, ICAL, HTML. Examples:
  - `http://www.bbc.co.uk/bbcone/programmes/schedules/london/today.JSON`,
  - `http://www.bbc.co.uk/radio3/programmes/schedules/2012/11/08.JSON`.
- Programmes data: one page per episode, brand, series, broadcast, version and clip for most TV and radio programmes in RDF, JSON, XML, HTML. Examples:
  - `http://www.bbc.co.uk/programmes/b01nqn80` (programme episode, html)
  - `http://www.bbc.co.uk/programmes/b01nqn7f.JSON` (version of that episode, containing some information about characters)
  - `http://www.bbc.co.uk/programmes/b0079t1p.rdf` (description of the brand Autumnwatch in RDF)
- External links to Wikipedia and DBpedia: in some cases Wikipedia and DBpedia links are included in these pages. Example: `http://www.bbc.co.uk/programmes/b015d4qz.JSON`. Note that the JSON is a sort of summary, the RDF has character data on version and links to DBpedia from the topic.

- MusicBrainz IDs and links: sometimes MusicBrainz data is present. Examples:
  - `http://www.bbc.co.uk/programmes/b01ngqsl` (html showing tracks played)
  - `http://www.bbc.co.uk/music/artists/59a1ef02-fd8b-4297-8fa5-875da633473b.rdf` (artist RDF showing links to MusicBrainz and Wikipedia)
  - `http://www.bbc.co.uk/music/artists/59a1ef02-fd8b-4297-8fa5-875da633473b.JSON` (similar as JSON)
  - `http://www.bbc.co.uk/1xtra/playlist/` (Radio 1 Xtra playlist) `http://www.bbc.co.uk/1xtra/playlist.JSON` (as JSON, showing the artist ids (MusicBrainz ids))
- Genres: genres are published as SKOS. Examples:
  - `http://www.bbc.co.uk/programmes/genres.rdf`
  - `http://www.bbc.co.uk/programmes/genres/childrens.rdf`.
- Topics pages, some linked to DBpedia, including people. Example:
  - `http://www.bbc.co.uk/programmes/topics.rdf` (list of topics)
  - `http://www.bbc.co.uk/programmes/topics/speech.rdf` (specific topic)
  - `http://www.bbc.co.uk/programmes/topics/stephen_fry.rdf`
- Specific data for some sub-sites:
  - `http://www.bbc.co.uk/nature/wildlife`
  - `http://www.bbc.co.uk/nature/life/Black_Grouse.rdf`
  - `http://www.bbc.co.uk/nature/life/Plant.rdf`
- Ontologies
  - Programmes Ontology: `http://purl.org/ontology/po/`
  - Sport ontology: `http://www.bbc.co.uk/ontologies/sport/`
  - wildlife ontology: `http://www.bbc.co.uk/ontologies/wildlife/`

  but also uses others, among them:
  - Music ontology: `http://purl.org/ontology/mo/`
  - FOAF: `http://xmlns.com/foaf/spec/`
  - Events ontology: `http://motools.sourceforge.net/event/event.html`

A complete list is available at `http://www.bbc.co.uk/programmes/developers`. The license under which the data is available is Attribution-Noncommercial-Share Alike Creative Commons License for BBC music and Creative Commons Attribution for BBC programmes.

2.2 Ratings data

IMDb Source for movie, TV and celebrity content. Information type:

- standard EPG information: director, writer, stars
- ratings with percentages and demographic information
- reviews both from users and critics. In the last case links to critics' reviews are provided.
- awards list
- miscellaneous links to other sources, $i.e.$ Wikipedia,

License: not Open.

API: no official API is available.

Rotten Tomatoes Rotten Tomatoes is a website devoted to reviews, information, and news of films, widely known as a film review aggregator. Information type:

- standard EPG information: director, writer, stars
- ratings with percentages and demographic information
- reviews both from users and critics.

All the content is provided in JSON format.

License: not Open

API: The Rotten Tomatoes API is a RESTful web service. The base URI to access all resources is `http://api.rottentomatoes.com/api/public/v1.0`. An APi key is required.

Facebook Open Graph Protocol Given the ID of the item, number of "likes" and number of "people talking about it" can be retrieved. Besides other standard descriptive information can be found, however these are inserted by the user who created the page, so they could unreliable. Facebook also contains a lot of information about shows and movies, often provided by their creators. In the future we intend to link this information to the ViSTA-TV data.

License: Open

API: HTTP request

Twitter Twitter makes available per item, whether this is a movie, a TV show or a person, number of tweets talking about it, and relative number of retweets.

License: Open

API:

- Timelines API: collections of tweets, ordered with the most recent first.
- tweets API
- Search API: find relevant tweets based on queries performed by your users.
- Friends & Followers: users follow their interests on Twitter through both one-way and mutual following relationships.
- Streaming API: aims at developers with data intensive needs. The streaming API allows for large quantities of keywords to be specified and tracked, retrieving geo-tagged tweets from a certain region, or have the public statuses of a user set returned. This requires to establish a long-lived HTTP connection and maintain that connection.
- Places & Geo API: allows to attach location data to tweets and discover tweets and locations.
- Trends API: allows to explore what's trending on Twitter.

2.3 Non-public data

This section describes the data need users permission in order to be retrieved.

Facebook

- personal information of the user: age, gender, living place, past living place(s), job, past job(s), interests, etc.
- activities, posts, likes
- friends information and activities

Twitter

- collection of the most recent tweets and retweets
- relationships of the authenticating user to the comma-separated list of up to 100 screen_names or user_ids provided. Values for connections can be: following, following_requested, followed_by, none.
- settings (including current trend, geo and sleep time information)
- the 20 most recent favorite tweets
- the information for saved searches

2.4 Summary

In this Section we presented an overview of the possible external sources we could use in the context of the project. The datasets selected from the LOD Cloud can be used immediately in the enrichment process, while less structured sources will need some pre-processing in order to be available for the enrichment process. Finally the non-public data will be available only when the application to be developed in WP5 will be in use.

3 Analysis of the raw data extracted from external sources

In this Section we will present some preliminary statistics about the selected datasets. We will start with LOD datasets, and continue with the other external sources.

3.1 LOD sources

Table **??** shows basic statistics about LOD datasets: number of objects, number of triples and how many in-links and out-links they already have.

Dataset	Objects	Triples	Out Links	In Links
DBpedia	3.77 Million	400 million	27.2 million	3.5 million
Freebase	23 million	337 million	3.9 million	3.4 million
BBC	139.236	60 million	43.237	0
BBC music	not available	20 million	23.000	903.435
NYT	10.467	345.889	23.400	20.289
MusicBrainz	not available	178 million	855.754	25.188
LinkedMDB	503.242	6 million	162.756	13800
GeoNames	8 million	94 million	0	117130
LinkedGeoData	1 billion	20 billion	53.204	53024

Table 2: Statistics about the selected LOD datasets

3.2 Ratings data

In this Section we provide a general analysis of the Ratings sources, although a lot of statistics cannot be provided because of the very general purpose of some sources.

IMDb IMDb has a database with more than 2 millions titles (movies and TV shows) and more than 5 million people. It has over 100 million unique users each month.

Rotten Tomatoes Rotten Tomatoes has a database with more than 250,000 titles and offers 850,000 review links. It has more than 7 millions users.

Facebook Open Graph Protocol Fig. 2 shows the typical type of websites Open Graph Protocol can be found on by distribution. We know of 40 categories that Open Graph Protocol websites can be found on[3].

Figure 2: Open Graph Protocol Top Ten Website Industry Distribution in the Top 100,000 Sites

Twitter Twitter is used by TV operators and broadcasters to try to encourage viewers to get involved. This can take the form of on-screen hashtags, having celebrities tweet while a show airs, or contests around a TV show. Notable examples include Comedy Central using Twitter hashtags during the Comedy Awards and BBC QuestionTime encouraging political debate. During the 2011 Oscars, there were over 10,000 tweets per minute-with the event racking up 1.8 million tweets overall. In the UK, daily television shows also receive attention on Twitter[4].

4 Using external sources for enrichment

Given description and analysis in the previous two sections, show how the selected sources can be used for the enrichment ADD EXPLANATION LIKING TO PREVIOUS SECTIONS

4.1 Enrichment example

In order to show the ability to enrich the standard EPG data, we worked on the following example. We used a program from BBC (`http://www.bbc.co.uk/programmes/b01nqn80`). The information from the EPG are:

- title: Autumnwatch
- episode title: Episode 3
- short synopsis: The team reveal the hidden world of the UK's nocturnal animals.
- medium synopsis: UK wildlife series. Chris Packham, Michaela Strachan and Martin Hughes-Games report back on the night's stake-out, revealing the hidden world of our nocturnal animals.
- long synopsis: The best stories and live wildlife action from Chris Packham, Michaela Strachan and Martin Hughes-Games, broadcasting from their base in the Scottish Highlands. The team report back on the night's stake-out, revealing the hidden world of our nocturnal animals. And new science uncovers how squirrels use deceit and theft to stockpile food for the tough times ahead.

[3] `http://trends.builtwith.com/docinfo/Open-Graph-Protocol`
[4] `http://en.wikipedia.org/wiki/Twitter_usage#In_television`

- genre: factual, science and nature
- format: documentaries

From Facebook we find that 1475 people liked this movie and theta 1055 people talked about it. From DBpedia we get:

- abstract from Wikipedia
- cast
- precededBy: a property that links a previous program similar to the actual one
- subject
- program creator

From NYTimes Search API we find 10 articles related with Scottish Highlands. We selected the ones that actually have the place mentioned in some of the field of the JSON reply. We find with 3 articles:

- "body" : ""Two Years at Sea", an enveloping portrait of a Scottish Highlands hermit from the experimental filmmaker Ben Rivers, is shot on 16-millimeter black-and-white film that feels at once tactile and evanescent. Similarly, the daily life we see of Jake Williams (the hermit who actually does live alone in a forest) is mundane and mysterious. The" , "byline" : "By NICOLAS RAPOLD" , "date" : "20121012" , "title" : "MOVIE REVIEW; Two Years at Sea" , "url" : "http:\/\/movies.nytimes.com\/2012\/10\/12\/movies\/two-years-at-sea-directed-by-ben-rivers.html"
- "body" : "LERWICK, SHETLAND – Across the Shetland Islands, an archipelago in the North Atlantic between the Scottish mainland and Norway, abandoned small stone houses are scattered around the windswept, almost treeless landscape. The ruined homes – a sight common in Scotland's islands, particularly in the Hebrides to the west of the mainland – are the" , "byline" : "By NICK FOSTER" , "date" : "20121005" , "title" : "In Remote and Rugged Shetland, a Dynamic Market" , "url" : "http:\/\/www.nytimes.com\/2012\/10\/05\/greathomesanddestinations\/05iht-reshetland05.html"
- "body" : "BIGFOOT hunters aren't the only ones out there searching for the unknown. Here are a few other "hunts" that travelers can join. The Atlantic Paranormal Society (413-478-3642; idealeventmanage.com). Perhaps the ghosting equivalent of the Bigfoot organization, this group organizes excursions to paranormal hot spots (its founders, Jason Hawes and" , "date" : "20120422" , "title" : "Search Parties: Ghosts, U.F.O.'s and Nessie" , "url" : "http:\/\/query.nytimes.com\/gst\/fullpage.html?res=9C02E2D71131F931A15757C0A9649D8B63"

Ratings information from IMDb:

This example shows that even if the starting EPG data does not provide much program metadata, we can retrieve the missing data and provide additional external data. Summarizing we provided: complete EPG data, ratings from 2 different sources, links to other programs and topics.

4.2 Enhancement of the features extracted by WP2

The external data service can enhance some of the feature extracted by WP2. In Table 4.2 we show the possible enrichment the external data service can provide.

Offline feature	Enrichment type
user:favorite_genre	URIs from LOD
user:favorite_person	URIs from LOD, newspaper articles, Facebook and Twitter info
user:favorite_program	URIs from LOD, newspaper articles, Facebook and Twitter info
EPG:synopsis	URIs from LOD, trend words from Twitter and Google

Table 3: WP2 extracted feature enhancement

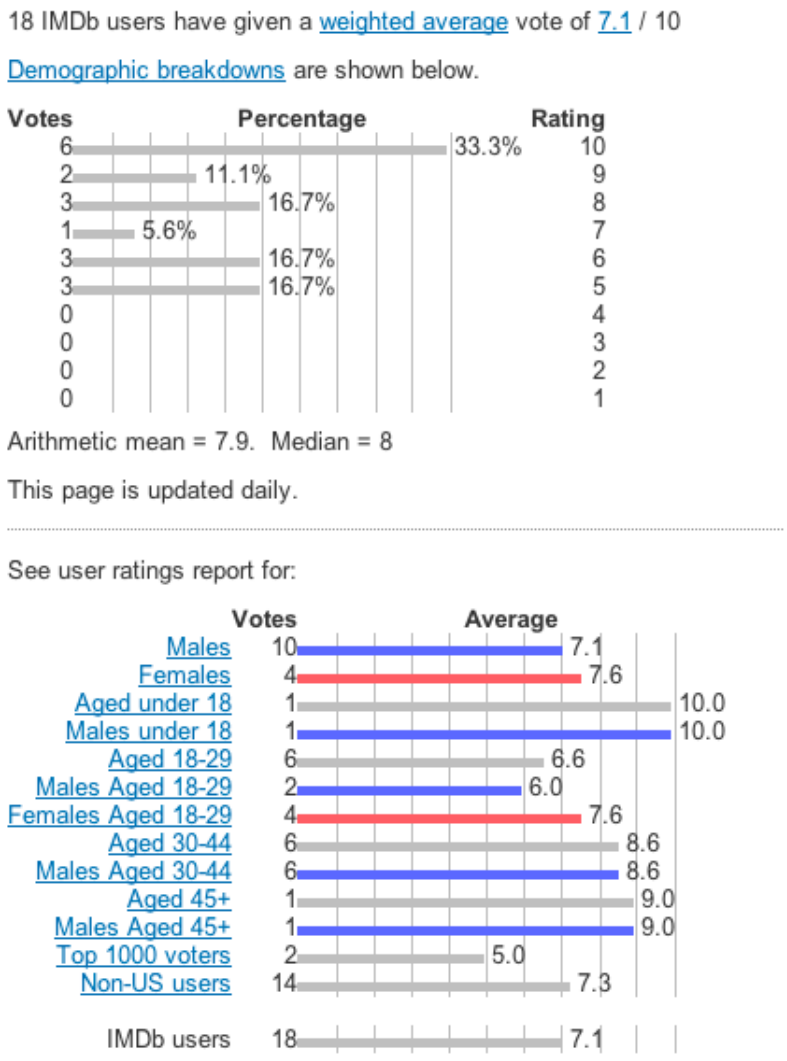

Figure 3: Screenshot from `http://www.imdb.com/title/tt0875050/ratings`

5 Design of data extraction workflows

The external data service will be performed offline, since the EPG data will be available at least one week in advance. The service will provide data for the next 7 days. This means that every day one day EPG enrichment will be performed. The enrichment of subtitles will be available after the first broadcast and will be available for the next one. The service will retrieve the EPG data directly from the providers (Zattoo and BBC). It will extract all the information needed to retrieve related external data, namely: title, cast, synopsis. The service will first perform a search in the LOD sources trying to match the named entities extracted with related URIs. For instance, given a title of a program and the related cast, we can find the program in one of the LOD sources, matching the title and the cast. For unstructured sources, the match is not straightforward, and some research needs to be done in order to identify possible heuristics to provide data about a specific program with a good level of confidence. Once the external data is collected, the service will make it available in the datastore.

A Licenses

In this appendix we describe the main types of licenses[5] we encountered during our analysis and how they allow to use the data in the context of the project, both with research and commercial uses.

A.1 Creative Common Attribution

This license allows to distribute, remix, tweak, and build upon other's work, even commercially, as long as you give credit for the original creation. This is the most accommodating of licenses Creative Commons license. See `http://creativecommons.org/licenses/by/3.0/` for the license deed.

A.2 Creative Commons Attribution Share-Alike

This license allows to remix, tweak, and build upon other's work even for commercial purposes, as long as you give credit and license to the new creations under the identical terms. This license is often compared to "copyleft" free and open source software licenses. All new works based on those work will carry the same license, so any derivatives will also allow commercial use. This is the license used by Wikipedia, and is recommended for materials that would benefit from incorporating content from Wikipedia and similarly licensed projects. See `http://creativecommons.org/licenses/by-sa/3.0/` for the license deed.

A.3 Attribution-Noncommercial-Share Alike Creative Commons Licence

This license allows to remix, tweak, and build upon other's work non-commercially, as long as you give credit and license to the new creations under the identical terms. See `http://creativecommons.org/licenses/by-nc-sa/3.0/` for the license deed.

[5] `http://creativecommons.org/licenses/`