NoTube: making the Web part of personalised TV

Balthasar Schopman
Vrije Universiteit
Amsterdam, Netherlands
schopman@cs.vu.nl

Dan Brickley
Vrije Universiteit
Amsterdam, Netherlands
danbri@cs.vu.nl

Vicky Buser
BBC
United Kingdom
vicky.buser1@bbc.co.uk

Libby Miller
BBC
United Kingdom
libby@nicecupoftea.org

ABSTRACT

With the steadily growing amount of TV channels, content selection has become a challenge to many users. While users have virtually no influence on what will be broadcasted on the majority of the TV channels, users have a great impact on what information is available on the World Wide Web. The NoTube project aims to close the gap between Web and TV by semantics.

Bits and pieces of personal and TV related data are scattered around the Web. NoTube aims to put the user back in the driver's seat by using data that is controlled by the user, *e.g.* from Facebook and Twitter, to recommend programmes that match the user's interests. By using the Linked Data cloud, semantic can be exploited to find complex relations between the user's interests and background information of programmes, resulting in potentially interesting recommendations.

Keywords

personalisation social networks linked data

1. INTRODUCTION

With the hundreds of channels available via modern TV providers, content selection and dealing with the vast amount of information become significant challenges for users. TV metadata is created and distributed by a small group of people, as a result of the closed-source information exchange protocols that are the standard for providing EPG data to users. The NoTube project puts the TV user back in the driver's seat by generating user profiles from data the user creates, *e.g.* tweets on Twitter,[1] and generates, *e.g.* a network of friends on Facebook,[2] and providing an interactive TV experience without information overload.

As much as the Web encourages users to actively participate, TV remains a largely passive experience. NoTube envisions a common ground for the Web and TV, where people actively contribute while watching TV, by discussing and recommending programmes. NoTube aims to make TV a community enhanced platform by overcoming the barriers of the physical location of the TV set, which makes TV a single-user (or single-family) experience.

People often have several clusters of personal data on the Web, such as their profiles on social networks and ratings on YouTube and IMDB. Analogously to the personal data, there are many isolated clusters of broadcast data on the Web, such as broadcast data on *electronic programme guides* (EPGs) and background knowledge on Wikipedia. The conjunction of all these bits and pieces of data would provide invaluable information of someone's interests and recommendations on TV broadcasts. Using user data that are already available online potentially solves the notorious *cold start problem*, *i.e.* that a recommendation system cannot provide useful recommendation until a substantial amount of statistical user activity data has been gathered. The NoTube vision is to realise a recommendation system that uses all these bits and pieces of data in conjunction. Semantic Web standards are important building blocks, as they enable the global identification mechanism of URIs and the means to define relations between data anywhere on the Web.

There are many content providing and playing applications, such as YouTube, BBC iPlayer, MythTV, Boxee, TiVo, Freevo and iTunes. Many of these interact only with their own proprietary resources and rarely make use of other existing services, even though many services are useful for data extraction or interaction with other applications. Exceptions are the BBC, who link their TV and radio metadata to Linked Data[3] such as MusicBrainz[4], and Boxee.[5] Boxee is a media centre that allows users to automatically publish their opinion on watched movies on Twitter, including automatically generated links to IMDB and iPlayer. NoTube does not aim to replace and outperform any of these existing applications, but to create a synergistic combination of these services by means of standardised APIs and protocols.

Due to its dynamic nature, the Web has been formed and shaped to the desires of its users. TV technology is not as dynamic as the Web and has kept its linear programming over the last decade, partially due to its business model where advertisements remain the most important source of income for commercial TV stations. The strong interaction between the desires of users and technology has had its im-

[1] http://twitter.com/
[2] http://facebook.com/

Web Science Conf. 2010, April 26-27, 2010, Raleigh, NC, USA.

[3] http://linkeddata.org/
[4] http://musicbrainz.org/
[5] http://boxee.tv/

pact on the Web and as the gap between the Web and TV experience grows, we aim to translate features of the Web to TV, such as the personalised and community-based aspects.

In this paper we discuss the aims of the NoTube project by describing three scenarios in Section 2 and the NoTube System architecture that realises these scenarios in Section 3. In Section 4 we describe usage scenarios that illustrate the discussed scenarios. In conclusion we discuss several challenges we face in Section 5 and state the progress of realising the scenarios in Section 6.

2. TV IN THE 21ST CENTURY

We illustrate three scenarios in the NoTube "Internet TV in the Social Web" use case.

Scenario 1: Recommendations based on Web behaviour. Jana wants to see recommendations of TV programmes based on her online social activity when she gets home at night. She talks a lot on Twitter and Facebook about what she reads and watches in the context of her online social life, and so do her friends. Jana does not see why she should have to explicitly tell any system what her preferences are, since a lot of data on her preferences are available online. She wants to see recommendations clearly featured on the user interface of her media centre.

Scenario 2: Do you want to know more?. While watching TV, Jana sometimes would like more information about a programme. She would like to be able to mark a programme to come back to it later and find out more about it. She does not necessarily want to have her laptop open all the time during this and neither does she want to interfere with the playing of the programme too much as she often watches TV with other people in the same room.

Scenario 3: Add to media centre's 'to watch' list, while away from home. While at work, Stephen sees an interesting documentary on an on-demand video service that he wants to watch tonight at home. He adds it to his NoTube bookmarks, so he can watch it at home on his media centre that evening.

In the following section we describe the system architecture, which identifies services that are needed to realise the scenarios illustrated above. Some of the necessary services already exist, such as the Linked Data cloud, media centres and remote control hardware devices. However, many open specifications and implementations of protocols and APIs need to be developed in order to realise these scenarios. Specifying these missing protocols and APIs will form a novel contribution to the TV media world, which NoTube aims to effectuate.

3. THE NOTUBE SYSTEM

Figure 1 shows the architecture of the NoTube System, which is a network of services.

Home Ambient. A home ambient consists of a display, a media centre and a remote control. The Home Ambient group forms the requirements of the home equipment a user requires to use the NoTube service.

Wider Home Ambient. The Wider Home Ambient group forms a large set of logically separate services that could all be deployed on hardware located at the user's premises. We discuss the distribution of services between the user's premises and the Web in Section 5.

Data warehouse. The group of services that is concerned with collecting and manipulating EPG data is called the data warehouse. The EPG data harvester may harvest more EPG data than the user has access to, so the data is filtered by the EPG availability filter component.

EPG data can be received via a digital cable or satellite connection in the DVB EIT standard. If not, it is harvested from other online resources like the popular XMLTV.[6] EPG metadata translated to the TV-Anytime[7] format, which is a rich TV metadata representation standard with a layered annotation approach. This layered annotation approach is similar to that of the BBC Programmes ontology.[8]

The programme data enrichment service performs named entity recognition and annotates plain text with semantic links, *i.e.* links to concepts in ontologies such as dbpedia.[9] By adding semantic links to the EPG data, the programmes can be placed in a broader context allowing for interesting links to be made between items. For example, in the context of two artists it may be derived that 'both artists worked in Detroit in the 1960's. The dbpedia relationship finder[10] shows many other examples of interesting relationships that can be deduced using Linked Data. An example where these relationship are used to find interesting relations between artists in the music domain is the Music Bore.[11]

Beancounter. The Beancounter fetches user activity data from the Web and generates user profiles. Mainly social network sites are used to collect user data, *e.g.* Facebook and Twitter. However, many other sites offer invaluable user activity streams. For example, LastFM lists a user's recently played songs and every song is annotated with MusicBrainz[12] identifiers. Since MusicBrainz is part of the Linked Data cloud, it is linked to the dbpedia concepts via an alignment service,[13] which enables access to potentially available background information of the user's music preferences.

The user profile is represented as a set of weighted interests,[14] where each interest is a concept in the NoTube cloud, which contains the Linked Data cloud, the MultimediaN cloud[15] and ontologies specific to the TV domain, such as TV-Anytime or BBC genres. An interest may be annotated with context, such as a location and a time interval, potentially enabling context-specific recommendations. For example, if someone always watches news in the afternoon, but never watches reruns of soap operas, the soap operas

[6] http://xmltv.org/
[7] http://www.etsi.org/WebSite/Technologies/TVAnytime.aspx
[8] http://www.bbc.co.uk/ontologies/programmes/
[9] http://dbpedia.org/
[10] http://relfinder.dbpedia.org/
[11] http://www.bbc.co.uk/blogs/radiolabs/2009/07/the_music_bore.shtml
[12] http://musicbrainz.org/
[13] http://sameas.org
[14] http://xmlns.notu.be/wi/
[15] http://eculture.cs.vu.nl/resources/datacloud/

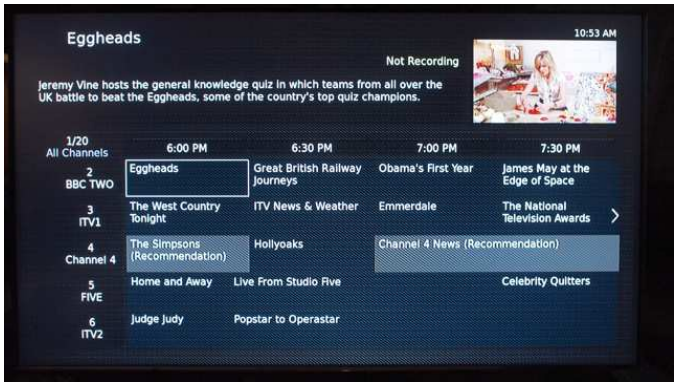

Figure 2: Screenshot of a media centre displaying the EPG with highlighted recommendations

that are broadcasted in the afternoon are not recommended to the user, but the one o'clock news is.

The user's activity on the home ambient generates a user activity stream that is recorded and stored. The Beancounter takes this user activity into account as well when generating the next user profile. Depending on the preferences of the user, the activity stream may be public or visible to friends. Visibility for friends enables real-time social recommendations, *i.e.* recommendations based on what your friends are watching.

Recommender. The recommender service is the nexus of the NoTube System, providing recommendations using the results of both the data warehouse and Beancounter groups. Since the user profile is a set of Semantic Web concepts, recommendations can be generated based on the average semantic distance [1] between the interests of the user and the potentially recommended item.

4. USAGE SCENARIOS

In this section we show usage scenarios corresponding to the three scenarios illustrated in Section 2.

The usage scenario of scenario 1, 2 and 3 are described in Table 1, 3 and 4 respectively. A screenshot of the media centre presenting the EPG with highlighted recommendations is shown in Figure 2.

Prior to scenario 1, Jana has used her Home computer to register herself for the NoTube Service and has added her Twitter and LastFM accounts. If Jana would have had a FOAF[16] profile when she registered, she could have referred to her FOAF data, instead of manually specifying her Twitter and LastFM usernames. Upon her registration the FOAF profile has been automatically created for future use. Directly after Jana's registration, the Beancounter group starts collecting user data in order to generate a user profile. This process is described as a usage scenario in Table 2.

5. DISCUSSION

Privacy. How do we address the challenge that people want personalisation without any compromise to their privacy?

[16] http://xmlns.com/foaf/spec/

The ability to share information online about programmes you have watched is already being made possible by experimental services such as 4iP's Test Tube Telly and Duncan Robertson's 'Radio Aunty' and Tellybox applications, which send details about what you are listening to and watching to the @radioandtvbot account on Twitter, if you opt in. Similarly Boxee allows you to match up your feed output to a twitter stream.

However, based on anecdotal evidence, it would seem that what people watch on TV is considered more private than the music they listen to, the DVDs they rent or the books they buy. People worry about how much time they can be seen to be wasting by watching trashy TV (indulging in their 'guilty pleasures'), which does not align with the online personas they have carefully cultivated. We therefore suspect that people are naturally quite guarded about sharing their TV viewing behaviours.

With NoTube Beancounters, there are additional privacy implications around the re-broadcasting of merged private and public data. Many people's intuitions about privacy are not applicable to new cases where their data is combined and re-broadcast, even if they have control over it, and particularly when they do not have a good understanding of the capabilities of software and the potential consequences of using it.

Consider the case where someone connects their TV to their private stash of illegally downloaded movies and also to twitter, and then twitter broadcasts the fact that they watched something they could not legally have watched with precise time and device information. There we have a new link created betweeen the person, the movie and the time, which provides evidence that they have done something illegal.

Other examples include the unwitting disclosure of personal information that people might prefer to keep private, such as medical conditions or political affiliations. Privacy settings could have complex and numerous consequences and it is difficult to predict all of them. MIT's Gaydar project[17] is a recent example of how personal information can be shared inadvertently.

We therefore need to ensure that sharing of data is a conscious choice for users. However it is difficult to support granular control of disclosure of personal information without imposing a heavy cognitive burden on the user.

Identifiers. Global channel and programme identifiers are crucial in a distributed environment. For example, users may share information about programmes, recommend programmes to each other or bookmark a broadcast from work to view in the evening (scenario 3 described in Section 2). In all cases users must be able to refer to a programme or broadcast with a location-independent identifier.

We have set several requirements for identifiers: identifiers are resolvable, type-based, content-based, human-readable and there should be a standard procedure to create or look up an identifier. To have resolvable identifiers, we can use URIs of online databases, such as IMDB, dbpedia or Freebase. However, these databases all lack certain requirements: IMDB and Freebase do not have human-readable URIs, but contain numerical identifiers. Dbpedia does not

[17] http://www.boston.com/bostonglobe/ideas/articles/2009/09/20/project_gaydar_an_mit_experiment_raises_new_questions_about_online_privacy/

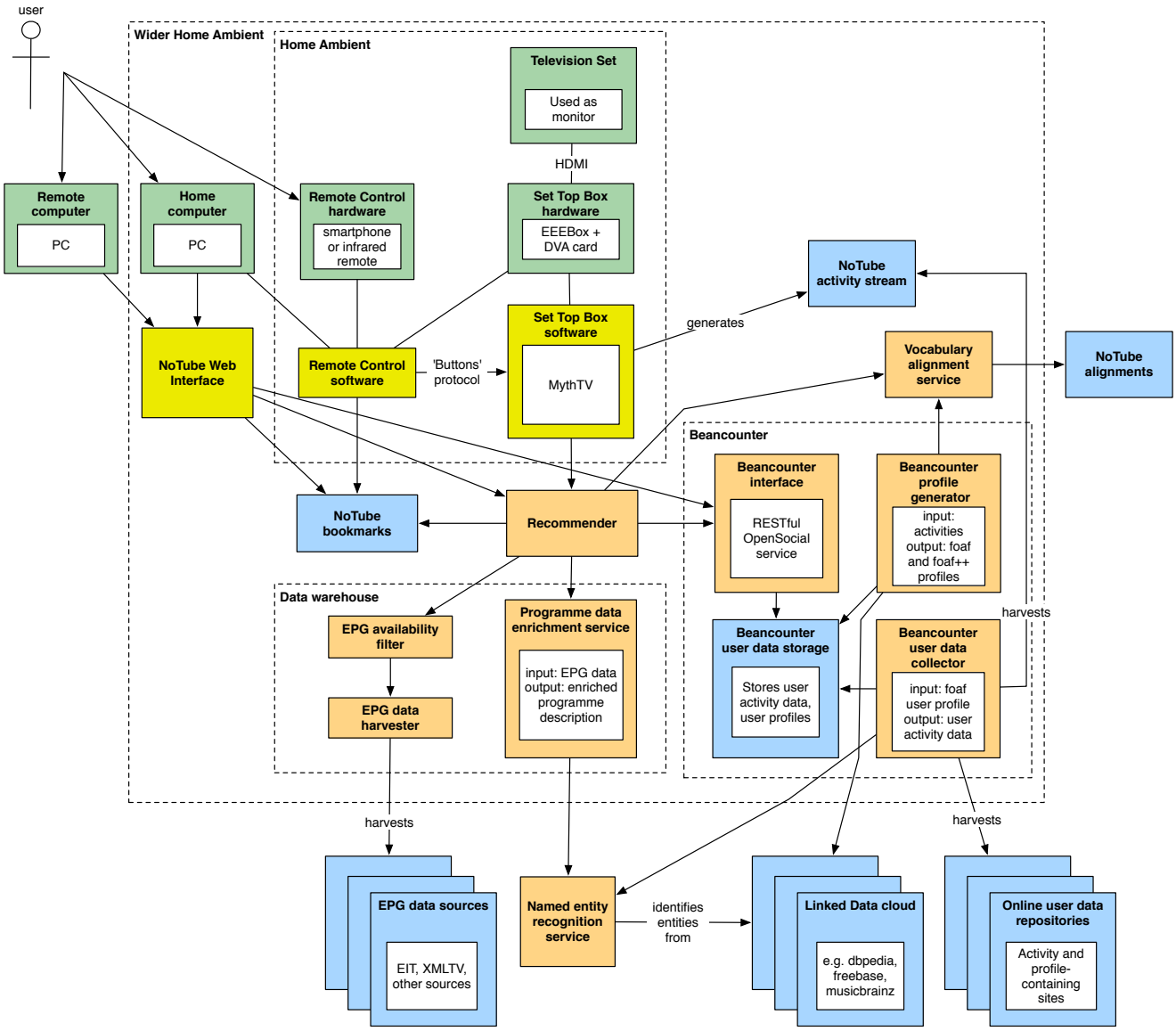

Figure 1: Architecture of the NoTube System. Boxes represent components providing services; links represent service calls, unless specified otherwise. Yellow and green services are respectively soft- and hardware services with which users have direct interaction. Blue services are interfaces to data repositories. Orange services provide algorithms that use mentioned data sources to enhance the TV experience.

Involved service(s)	Action	Remark
set top box hardware, television set	Jana turns on the TV and media centre.	
set top box hardware, set top box software, remote control hardware	service discovery phase: set top box detects presence of Jana's smartphone	
set top box hardware, set top box software, remote control hardware	authentication protocol	Jana has paired the set top box with her smartphone and enabled the smartphone presence authentication protocol, so she does not have to enter her username and password to start using her Home Ambient
set top box software, Recommender	set top box software requests recommendations from Recommender	request contains user authentication data and list of available TV channels
Recommender, Beancounter interface	Recommender requests user profile	
Beancounter interface, Beancounter user data storage	Beancounter interface fetches user profile from Beancounter user data storage	
Beancounter interface, Recommender	Beancounter interface sends user profile to Recommender	
Recommender, EPG availability filter	Recommender requests filtered EPG data	filtered data contains only EPG data of channels available to the user
EPG availability filter, EPG data harvester	EPG availability filter requests all available EPG data	
EPG data harvester, EPG availability filter	EPG data harvester fetches all available EPG data and sends it to the EPG availability filter	EPG data might be harvested beforehand and cached
EPG availability filter	EPG data filtering process, all EPG data of channels the user cannot access on his/her set top box is purged	
EPG availability filter, Recommender	EPG availability filter sends filtered EPG data to Recommender	
Recommender, programme data enrichment service	Recommender sends request to enrich EPG data	
programme data enrichment service, Named entity recognition service	EPG data is enriched by adding links to semantic concepts to recognised entities	optionally background data, *i.e.* associated concepts, are added
programme data enrichment service, Recommender	programme data enrichment service sends enriched EPG data to Recommender	
Recommender, Vocabulary alignment service	Recommender generates content-based recommendations based on weighted interests in user profile and enriched EPG data	Vocabulary alignment service is used to match identical concepts of different vocabularies, *e.g.* genres of the BBC and IMDB genre vocabularies
Recommender, set top box software	Recommender sends recommendations to set top box software	
set top box software, set top box hardware, Television set	recommendations are presented to the user	

Table 1: Usage scenario 1: Recommendations based on Web behaviour

Involved service(s)	Action	Remark
Beancounter user data collector, Online user data repositories	fetch user data of accounts in FOAF profile from Online user data repositories	Start of user data collection process. Input: FOAF profile that specifies user's social network identities.
Beancounter user data collector, named entity recognition service, Linked Data cloud	Named entity recognition service enriches user data with links to concepts in Linked Data cloud.	
Beancounter user data collector, Beancounter user data storage	store enriched user activity data	
Beancounter profile generator, Beancounter user data collector	fetch enriched user activity data	Start of user profile generation process.
Beancounter profile generator, Linked Data cloud	Background data, *i.e.* links to associated concepts, are added to each item in the user activity data using relations in the Linked Data cloud.	
Beancounter profile generator, Vocabulary alignment service	User profile is generated by counting occurrences of concepts in the user activity data. The Vocabulary alignment service is used to aggregate identical concepts of different vocabularies.	The result is a set of weighted interests, interests are semantic concepts in NoTube cloud.
Beancounter profile generator, Beancounter user data storage	Generated user profile is stored in Beancounter user data storage	

Table 2: Beancounter user data collection and user profile generation

Involved service(s)	Action	Remark
Remote Control Hardware		Jana watches a TV programme and sees something she would like to know more about. She picks up the remote control, *i.e.* her smartphone, in order to press the 'I want to know more' button.
Remote Control Software, Set top box software	Remote Control Software requests data of currently watched programme from Set top box software.	
Set top box software, Recommender	Set top box software requests data of currently watched programme from Recommender.	
Recommender, Set top box software	Recommender sends cached programme data to Set top box software. If cached data is unavailable, it is acquired as during the recommendation process as described Table 1.	Recommender has enriched programme data in cache from generating recommendations process.
Set top box software, Remote Control Software	Set top box software sends data of currently watched programme to Remote Control Software.	Programme data sent to remote control is gathered by data from various sources, *e.g.* from both the original EPG data, Wikipedia and Freebase
Remote Control Software	Remote Control Software displays programme information.	Jana watches the programme data displayed on her smartphone without interrupting the broadcast that is being displayed on the TV set.

Table 3: Scenario 2: the 'I want to know more' button

Involved service(s)	Action	Remark
Remote computer, NoTube Web Interface		Steven is looking at the EPG of tonight and sees a broadcast he would like to watch. He clicks the 'watch this tonight' button.
Remote computer, NoTube Web Interface, NoTube bookmarks	NoTube Web Interface adds this program to NoTube bookmarks.	
NoTube bookmarks	Broadcast is bookmarked.	When the set top box is turned on tonight, the Recommender will retrieve bookmarks and immediately recommend this broadcast

Table 4: Scenario 3: add to media centre's 'to watch' list while away from home

incorporate changes in Wikipedia real-time, resulting in a long propagation time from a Wikipedia update the corresponding dbpedia update and poor coverage of recent TV shows.

The human-readable, type- and content-based requirements imply that the URIs referencing different types (*e.g.* channel, broadcast) should be distinguishable and it should be possible to understand what entity it identifies. For example, the identifier of the TV channel 'RAI uno' could be http://purl.org/identifiers/channel/RAI1

Distribution of services. As hinted in Section 3 the Wider Home Ambient group shown is Figure 1 forms a large set of logically separate services that could all be deployed on hardware located at the user's premises. It remains an open question which distribution of service deployment is optimal, *i.e.* do we favor any service to be deployed on the user's premises or on the Web?

There are several trade-offs to be considered, such as (the user's sense of) security and privacy versus maintainability and availability of the service. We are seeing a shift from both software and data on the home computer to the cloud, where Google seems a pioneer of Web applications with Google Docs[18] and the upcoming Chromium OS.[19] Advantages of a service such as Google Docs are that the user is not bothered with keeping the software up to date and one can access his/her documents from any computer with an Internet connection. However, storing documents online seems a disadvantage to many users, because it is considered insecure to store your data on the Web.

As connectivity increases and we see a shift towards Web services and storing our data on the Web, will TV in the 21^{st} century become a Web service or are users more comfortable with their data locally stored on a set top box?

Other challenges include:

- Managing time and geographical restrictions that are common to TV content.
- TV is a world dominated by closed, commercial standards and copyright restrictions. Open source media centres are rare and the best EPG data are harvested from TV guide websites.
- Gaining access to protected data in a secure fashion. Often services ask users to give their username and password, which is a serious risk hazard. OAuth[20] and OpenID[21] are fairly new and unadopted standards that try to solve this.

6. PROGRESS

The NoTube project started in February 2009, *i.e.* one year ago at the time of writing this paper. Some services we describe in this paper are implemented, while others are partially implemented or still in design.

A non-exhaustive list of services that are developed for the NoTube project:

- Beancounter :
  - user profile generator.
  - OpenSocial[22] interface.
  - last.fm[23] and Glue[24] user data collectors
- Alignments between BBC genres, TV-Anytime and YouTube genres.
- Named entity recognition service LUPedia.[25]
- Mocked up demonstrator: the iZapper.[26]

Future work within the NoTube project will include:

- TV bookmarks.
- Support for FOAF profiles in the Beancounter user input.
- A content-based recommender service that uses the Linked Data cloud and contextual data, *e.g.* temporal data.
- Service discovery functionality using the XMPP discovery protocol,[27] to have the smartphone and set top box discover each other's presence as described in Table 1.

[18] http://docs.google.com/
[19] http://www.chromium.org/chromium-os
[20] http://oauth.net/
[21] http://openid.net/
[22] http://code.google.com/apis/opensocial/
[23] http://last.fm/
[24] http://getglue.com/
[25] http://lupedia.ontotext.com/
[26] http://services.notube.tv/
[27] http://xmpp.org/extensions/xep-0030.html

7. ACKNOWLEDGEMENTS

This work is supported by the EU IP Project NoTube.

8. ADDITIONAL AUTHORS

Chris van Aart (Vrije Universiteit Amsterdam, email: chris@vanaart.com),
Lora Aroyo (Vrije Universiteit Amsterdam, email: l.m.aroyo@cs.vu.nl),
Stefano Linguerri (Pro-netics, Rome, email: stefano.linguerri@gmail.com),
Véronique Malaisé (Vrije Universiteit Amsterdam, email: vmalaise@few.vu.nl),
Michele Minno (Pro-netics, Rome, email: michele.minno@gmail.com),
Michele Mostarda (Pro-netics, Rome, email: michele@asemantics.com),
Lyndon Nixon (STI International, email: lyndon.nixon@sti2.org),
Davide Palmisano (Pro-netics, Rome, email: davide@asemantics.com),
Yves Raimond (BBC UK, email: yves.raimond@gmail.com),
Ronald Siebes (Vrije Universiteit Amsterdam, email: rm.siebes@few.vu.nl)

9. REFERENCES

[1] Yiwen Wang, Natalia Stash, Lora Aroyo, Peter Gorgels, Lloyd Rutledge, and Guus Schreiber. Recommendations based on semantically enriched museum collections. *Web Semant.*, 6(4):283–290, 2008.