Non-Intrusive User Interfaces for Interactive Digital Television Experiences

Pablo Cesar, Dick C. A. Bulterman, Zeljko Obrenovic, Julien Ducret, Samuel Cruz-Lara

Non-Intrusive User Interfaces for Interactive Digital Television Experiences

Pablo Cesar[1], Dick C.A. Bulterman[1], Zeljko Obrenovic[1], Julien Ducret[2], and Samuel Cruz-Lara[2]

[1] CWI: Centrum voor Wiskunde en Informatica
Kruislaan 413
1098 SJ Amsterdam, The Netherlands
p.s.cesar@cwi.nl, zeljko.obrenovic@cwi.nl, dick.bulterman@cwi.nl,
[2] LORIA / INRIA Lorraine
Campus Scientifique - BP 239
54506 Vandoeuvre-lès-Nancy, France
Samuel.Cruz-Lara@loria.fr, Julien.Ducret@loria.fr

Abstract. This paper presents a model and architecture for non-intrusive user interfaces in the interactive digital TV domain. The model is based on two concepts: non-monolithic rendering for content consumption and actions descriptions for user interaction. In the first case, subsets of the multimedia content can be delivered to different rendering components (e.g., video to the TV screen and extra information to a handheld device). In the second case, we differentiate between actions, handlers, and activators. An action is the description of the user intentions, a handler implements that action, and an activator is the user interface of the action. Because we define actions instead of user interfaces, the implementation of the activators can take multiple forms: conventional user interfaces (using gestures or speech) and intelligent interfaces, in which the actions are derived from a set of parameters (e.g., number of people in the room or distance to the TV).

1 Introduction

Watching television is usually a shared experience. The family or a group of friends sit in front of a shared output device (screen) and the interaction is carried out by using a single shared input device (remote control). In this paper, we propose a model, in which users can use personal devices (e.g., handheld) and sensory enhanced everyday objects to consume and interact with television content. We refer to this model as non-intrusive user interfaces because the personal activities do not interrupt the shared experience, but enrich the individual experience.

In our research we focus on the last stage of the media distribution chain; when the user is actually consuming and interacting with the digital content. In order to keep the paper in scope, we assume that the user has already selected a piece of content, probably with the help of a recommender system and that the

content (or a link to the content) is stored on a local home media server such as a Personal Digital Recorder (PDR) with different network interfaces (e.g., P2P network).

2 Related Work

The main motivation of our research is to provide the user, or a group of users, advanced control over the consumed content. We share the view presented by Baker [1] that current intrusive interfaces are not the solution. The future of interactive television is to provide a valued experience and not only just services. This effective group experience should rely on non-intrusive user interfaces for content selection, navigation, rendering, and interaction. Moreover, the source material does not have to be limited to the broadcasted content, but in the Web2.0 era a more convergent approach should be taken.

Much of the research on content selection within a digital television environment has focused on the macro-level concerns of selecting an entire program among a wide range of content available to a user. This is often done by some form of recommender system [2]. While we agree that recommender systems will play an important role in the future, they provide little or no assistance in navigating through content once it arrives in the home.

Micro-level content selection is often required if content-based navigation through a program is to be supported. Finding a program that fits the profile of a user is useful, but being able to navigate or select stories within that program is equally essential. A micro-level recommender system is required that processes the base content and then selects fragments of interest. (This is not an exclusive selections, but could be used in a customized navigation control interface).

For micro-level content selection, the most relevant standard today is the TV-Anytime Forum[3]. Interesting research in this area includes the UP-TV project. The UP-TV project [3] has presented a program guide that can be controlled and managed (e.g., delete programs) from personal devices (e.g., handheld devices). Our research work, though, focuses on a finer level of granularity; on how fragments within a program can be managed and customized using a variety of end-devices.

The Current standardised declarative solutions for interactive digital television are Digital Video Broadcasting - HyperText Markup Language (DVB-HTML)[4] in Europe, Advanced Common Application Platform - X (ACAP-X)[5] in USA, and Broadcast Markup Language (BML)[6] in Japan. These solutions are based on a number of eXtensible Markup Language (XHTML) modules and other World Wide Web Consortium (W3C) standards such as Cascading Style Sheets (CSS) and Document Object Model (DOM). So, they force to use a non-declarative solution, such as ECMAScript or Java, for modeling the temporal

[3] http://www.tv-anytime.org
[4] http://www.mhp.org/
[5] http://www.atsc.org/
[6] http://www.arib.or.jp/english/index.html

relationship between media elements in the document. In this paper, we use more promising current solutions instead such as the Brazilian middleware standard [4], called Nested Context Language (NCL), and SMIL next generation [5], as proposed by W3C.

Regarding content rendering and interaction, Jensen [6] defines three basic types of interactive television: enhanced (e.g., teletext) personalized (e.g., pause/play content stream using a PDR), and complete interactive (i.e., return channel). In this paper, we extend Jensen's categorization with a new television paradigm: viewer-side content enrichment. In this paradigm, the viewer is transformed into an active agent, exercising more direct control over content consumption, creation and sharing. A key element of our paradigm is that, unlike with the PC, the television viewer remains essentially a content consumer who participates in an ambient process of incremental content editing. Similar results, but intended to broadcasters instead of end-users has been presented by Costa [7].

Finally, Chorianopoulos argues that traditional metaphors cannot be applied to digital television [8]. He proposes a metaphor called the Virtual Channel: dynamic synthesis of discrete video, graphics, and data control at the consumer's digital set-top box. In this paper we extend that notion by providing a system that can retrieve enriched content from external web services such as Wikipedia.

3 Contribution

The main contribution of this paper is a model and an architecture that provides the user an enhanced experience in relation to traditional interactive television services such as the red button or SMS voting solutions. We can divide this contribution into three different categories: content modeling, content consumption, and user interaction.

First, we propose to model the content using a rich-description standard such as SMIL [9] in combination with TV-Anytime metadata descriptions. The major benefit of this approach is that the content, and fragments of the content, can be enriched by different parties at different times. For example, content creators can include enriched material at the creation stage, while individuals might further enhance the content at viewing time. Moreover, at viewing time, content enrichment can be obtained from different freely available resources such as Wikipedia.

Second, we study the differences between the private and the share space. For example, the television in the living room is a shared space between family members, while a handheld device is a private space. This paper proposes as an innovation the development of a non-monolithic multimedia player that is capable of rendering parts of content into different output devices depending on the share/private nature of the content.

Finally, we propose a model for user interaction based on actions instead of interfaces [10–12]. We define three components: actions, handlers, and activators. The action is the description of the user intention (e.g., pause content or add

media), the handler is the implementation of the action, and the activator is the user interface for the action (e.g., play button, speech recognition engine, or gesture). By designing actions, the actual activators can be implemented in a variety of ways (e.g., gestures, voice, or sensory enhanced everyday objects). The major benefit of this solution is that the user is not limited to the remote control interaction, but can use his personal device, or even enhanced everyday objects can be used to interact with the content.

4 Architecture

This section introduces the architecture proposed in this paper for providing non-intrusive user interfaces in the home environment. Figure 1 shows the architecture of our system, that includes the following components:

- an intelligent and flexible middleware component, called AMICO
- a non-monolithic SMIL rendering component, the Ambulant Player [13]
- the actions handler called Ambulant Annotator [14]

The next subsections describe in more detail the components of our architecture. First, we describe the how to model interactive digital television content for providing the users an enriched experience. Second, we introduce the AMICO middleware. Then, we indicate the extensions provided to the Ambulant Player and Annotator components. Finally, the interfaces between the components are explained together with some examples of usage.

4.1 Content Modeling

This architecture makes use of an enriched description of the multimedia content that includes SMIL files linked to TV-Anytime metadata description and Multi Lingual Information Framework (MLIF) [15, 16] textual content. Moreover, external services such as BabelFish, Flickr, YouTube, and Wikipedia provide additional content. Finally, a number of activators and rendering components can be utilized to consume and interact with the content. Some examples include the television set, handheld devices, and video cameras.

As introduced in the related work, television content can be modeled using a number of different standards. Instead of the text-centric solutions common to the web (e.g., XHTML), we propose to use SMIL; a media-centric solution in which the spatial and temporal synchronization between the media elements are described as high-level constructs. This solution provides a number of benefits: SMIL code is small, it is easily verifiable, it allows content associations to be defined easily, it provides a separation between logical and physical content, and it provides as base for license-free implementation on a wide range of platforms.

Our system includes support for MLIF. MLIF should be considered as a unified conceptual representation of multilingual content and its related segmentation (i.e. linguistic granularity). MLIF is being designed with the objective of

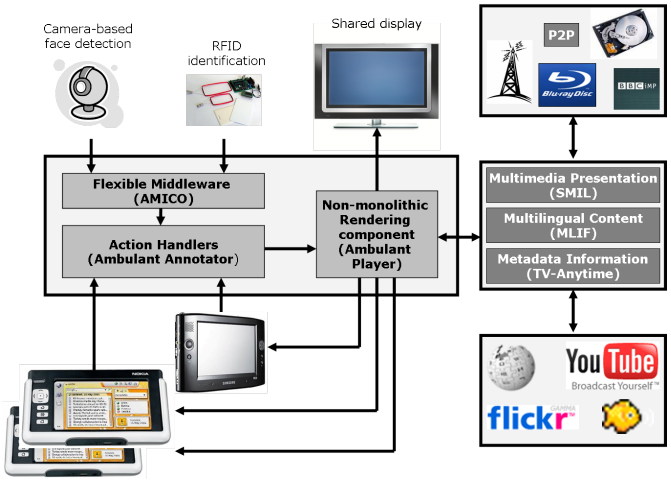

Fig. 1. System Software of the Proposed Architecture.

providing a common conceptual model and a platform allowing interoperability among several translation and localization standards, and by extension, their committed tools. The asset of MLIF is the interoperability which allows experts to gather, under the same conceptual unit, various tools and representations related to multilingual data. In addition, MLIF will also make it possible to evaluate and to compare these multilingual resources and tools.

Being able to handle and, for a user, being able to interact with digital media in order to deal with subtitles or to retrieve some complementary multilingual textual information, is not enough. Within MLIF, besides multilinguality-related issues, one of the most important innovations we are proposing is to be able to deal with different hierarchies of textual segments: linguistic granularity (i.e. sentences, words, syllables), document structure (i.e. title, paragraph, section), or any other personalized textual segmentation which may allow, for example, to associate time and format to any specific segment.

4.2 The Brokering Infrastructure: AMICO

Supporting novel interaction modalities with TV requires usage of many heterogeneous software modules, such as sensors, reasoning tools, and web services. Desired functionality is often available in a form of open-source and free software. The open-source community has developed a number of freely available components can provide as additional interaction modalities. Examples include libraries for vision-based interaction modalities, lexical tools, and speech input and output for many languages. The main problem in usage of these components is that they are developed for other purposes, in diverse implementation environ-

ments, following standards and conventions often incompatible with multimedia and TV standards.

To solve some of these problems we have developed Adaptable Multi-Interface Communicator (AMICO), an infrastructure that facilitates efficient reuse and integration of heterogeneous software components and services. The main contribution of AMICO is in enabling the syntactic and semantic interoperability between a variety of integration mechanisms used by heterogeneous components. The infrastructure and some examples are available at the SourceForge web site [7]. AMICO is realised as a Java application, and has been tested on several operating systems.

The proposed brokering infrastructure is based on the publish-subscribe design pattern. It is well suited for integration of loosely-coupled parties, and often used in context-aware and collaborative computing. A publisher may update a shared data repository without being concerned with whether any subscribers are listening for updates. When using simple data structures, the loosely coupled approach can be highly adaptable, so that new applications can both reuse existing data in the repository and add their own data without breaking the infrastructure. This approach is also fault tolerant, as components run as independent processes. In the loosely coupled model, components can run on different machines in a distributed environment. Components communicate by exchanging events through a shared data repository consisting of named slots. Components can update the slots, and register for notifications about changes. A key difference between our infrastructure and regular notification services is based on our requirement for supporting more than one integration interface. AMICO provides a unified view on different communication interfaces, based on a common space to interconnect them. It supports several widely used standard communication protocols. AMICO is extensible, and it is possible to add new communication interfaces. Most of the communication adapters are bidirectional. For example, an XML-RPC communication interface may run an XMLRPC server, enabling other modules to update and read data through this interface, and it also enables the definition of XML-RPC adapters that map this data to parameters of method calls on other XML-RPC servers. It is also possible to directly communicate with AMICO using a TCP connection, or by sending UDP packages.

4.3 Non-Monolithic Rendering of Content and Actions Handlers: Ambulant

In previous work, we have described the first prototype implementations of the Ambulant player and annotator [14]. The player is a multimedia rendering environment that supports SMIL 2.1, while the annotator is an extension of the player that is a DOM-like interfaces to and from the player implemented in Python. Together, player and annotator, provides viewer-side enrichment of multimedia content functionality at viewing time.

[7] http://amico.sourceforge.net

In addition to those capabilities, this paper introduces two innovative extensions to Ambulant:

- end-user actions handler: the Ambulant Annotator handles the user actions. These actions can come from personal activators (e.g., Nokia770) or from AMICO middleware. Some simple actions the annotator understands are play/pause; more complex actions include, for example, provide me extra information in French about the movie I am watching now.
- non-monolithic rendering: the Ambulant Player is responsible of targeting different parts of the presentation and content to different rendering devices. For example, the Ambulant Player can render extra information or commercials in my personal device.

4.4 Interfaces

This subsection describes the actual interfaces and responsabilities of the components of the architecture.

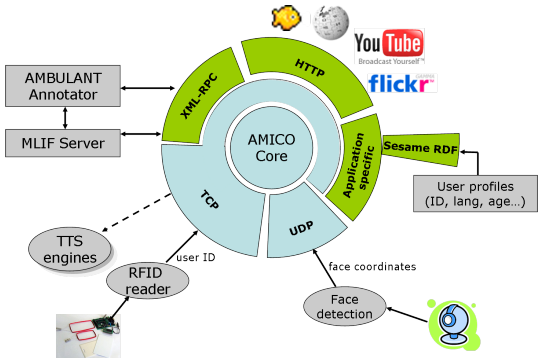

Fig. 2. AMICO Interfaces.

Figure 2 shows the interfaces of the AMICO middleware. As mentioned before, AMICO is an infrastructure that facilitates reuse and integration of components. AMICO provides the following functionality:

- Non-intrusive activators: catches, handles, and interprets the input from non-intrusive activators such as the video camera and the RFID reader.
- User Profile: retrieves and utilises the different user profiles encoded in Resource Description Framework (RDF) files.
- External Services: might retrieve content from external services such as Wikipedia.
- Output Transformation: provides output transformation features such as text-to-speech using a dedicate engine and on line translation using BabelFish.

Figure 3 shows the interfaces of the Ambulant Player and Annotator. It provides the following functionality

- Action Handling: the Ambulant Annotator handles the user input. The actions might come directly from the activators or from the AMICO middleware.
- Content Retrieval: the Ambulant Player has the capability of accessing different content resources. These resources include broadcast, optical disk, P2P network, and local storage. In addition, the Ambulant Player might request content from external services via the AMICO middleware.
- Non-monolithic Rendering: the Ambulant Player can divide and target the multimedia presentation to different rendering components. These components include the television set and other personal devices.

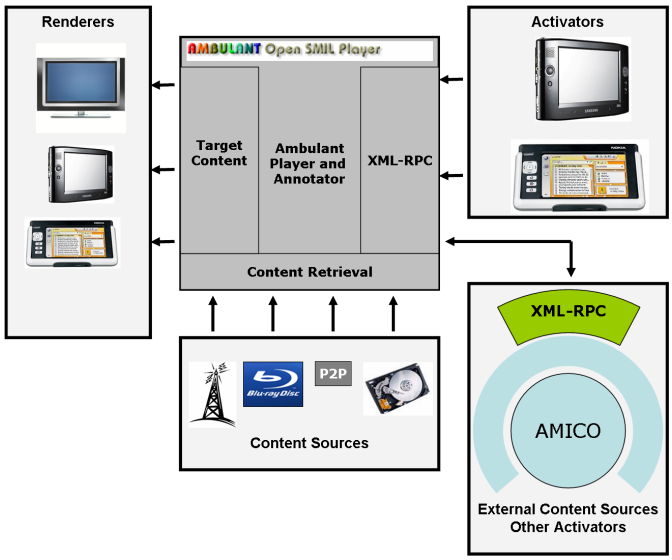

Fig. 3. Ambulant Player and Annotator Interfaces.

5 Scenario

In order to validate the ideas presented in this paper, this section presents an implemented scenario together with an analysis of the benefits of our solution over traditional interactive digital television systems.

The main actor of our scenarios is called Dick, a USA citizen. He has recently moved to the Netherlands with his wife, a Dutch woman. The scenario is divided

into three parts: non-monolithic rendering, non-intrusive input, and rich user interaction capabilities.

Non-Monolithic rendering: Dick and his wife are sitting in front to the television screen. His Personal device (e.g., Nokia 770) acts as an extended remote control with rendering capabilities. His personal device registers to the Ambulant Annotator. On the one hand, Dick can navigate the media content, as shown in Figures 4(a) and 4(b). On the other hand, because of the non-monolithic nature of the player, the personal devices are informed when interesting extra fragments are available for Dick. In both cases the content is rendered in the personal device and, thus, do not disturb the shared experience. The personal content might include, for example, instant translation of sentences he might not yet understand in Dutch, personalised commercials, or extra features extracted from web services.

(a) Packages Navigation. (b) Programs Navigation.

Fig. 4. Screenshot of Content Navigation in a Handheld Device.

Non-Intrusive input: in this scenario non-intrusive activators are used to interact with the television content. For example, sensors with RFID readers and camera-based face detectors can be used to detect the users identity and their distance to the television set. This data can be used, for example, to identify the common language of all the users, which then can be used to automatically select the subtitle language for the shared display. The data gathered from the camera-based detector, can be used, for example, to enable context-sensitive playback, where the player can be paused when there is no one in front of the screen, or warning message presented if the user is too close to the screen. To enable flexible content adaptation, we have developed a middleware that can enable customized mapping between sensing and presentation components, and derivation of actions from lower-level sensor values. Other important scneario will be non-intrusive health-care applications.

User Interaction Capabilities: in addition of providing a shared experience at home, our system provides a shared experience for connected people. For example, Figure 5(a) shows the content Dick is watching on the television screen. At some moment, he uses the Ambulant Annotator to enrich the content.

Figure 5(b) shows the interface in his handheld device. This enriched content is then shared with the, for example, his brother living in the USA using, for example, a P2P network.

(a) Content in the TV Screen.

(b) Enrichment Interface in a Handheld Device.

Fig. 5. Non-Intrusive Rendering.

This scenario show the two main contributions of this paper: non-monolithic rendering of content and non-intrusive user input. Based on a rich television content model, we believe they are the cornerstone of valued group experiences. Clear advantages of our system over current solutions include the capacity of targeting the personal content to where it belongs: to personal devices; the possibility of linking media content in packages or experiences, and the support for a variety of input mechanisms due to the action descriptions. For example, in addition to the ones mentioned earlier, AMICO supports voice input and, even more interestengly, an intelligent pillow interface for controlling the media playback.

6 Conclusion and Future Work

The most relevant conclusion of this paper is that the future of interactive digital television lies in providing the users a rich experience and not in providing only appealing services. Based on that conclusion, this paper has presented a model and architecture for non-intrusive user interfaces in the home environment. This model takes into account the differences between the share space (e.g., television set) and the private space at home (e.g., handheld device). The main contribution of this paper is the proposal of an architecture based on non-monolithic rendering of content and device-independent user interaction. In the first case, the architecture provides the mechanisms to target specific parts of the digital content at home to different rendering components (e.g., high-definition content to the television set and personal material to handheld devices). This way, the personal experience of the user is enriched, while the share experience is not

disturbed. In the second case, user interaction is not limited to the intrusive remote control paradigm. Even though in some cases such interaction is desirable, we propose to enrich the user potential impact on the content. Some examples include the use of personal devices for personal content interaction. In addition, other devices such as personal identifiers and a camera can register the identity and context of the user in a non-intrusive manner. Based on those variables, our system can derive actions on the multimedia content (e.g., to pause the show when there is nobody in front of the TV).

In addition to non-monolithic rendering of digital content and descriptions of the actions, this paper presents a solution for modeling interactive digital television content. The key question that this paper handles is how to model interactive television content in a rich and scalable manner. The solution provided by this paper is to use SMIL language linked to TV-Anytime metadata and to MLIF multi-lingual content. The major advantages of this solution is that the content can be fine-grained annotated, other resources than the broadcasted content can be included in the television packages, and further enrichments can be provided by professional and amateur users. First, we can include metadata content at the package, program, and segments levels. Second, because of the nature of SMIL, we can include resources coming from different sources in each television package. Thus, each television package will become a group experience rather than a service. Finally, because SMIL is an open standard, in addition to content producers other parties can enrich the television package. In the business level, we can think for example of broadcast companies including personalized commercials in the television package. In the social level, we can think of pal-users enriching the television material with their personal insertions. The solution we propose in this paper, SMIL, allows to further include different layers of content and metadata insertions.

Finally, in order to validate the ideas presented in this paper we present an implemented prototype of the architecture and use it in a scenario. This scenario is based on two of the most important research issues in interactive digital television: social television and ambient technology. Based on the scenario, we can conclude that our solution extends current interactive television systems by providing non-intrusive user interfaces. Future work includes to, first, describe business models based on the ideas presented in this paper, and, second, carry out a number of user studies.

7 Acknowledgements

This work was supported by the ITEA project Passepartout, by the NWO project BRICKS, and the IST-FP6 project SPICE. The development of Ambulant is supported by NLnet.

References

1. Baker, K.: Intrusive interactivity is not an ambient experience. IEEE Multimedia **13** (2006) 4–7

2. Blanco, Y., Pazos, J.J., Gil, A., Ramos, M., Fernández, A., Díaz, R.P., López, M., Barragáns, B.: AVATAR: an approach based on semantic reasoning to recommend personalized tv programs. In: Special interest tracks and posters of the 14th international conference on World Wide Web. (2005) 1078–1079 ISBN 1-59593-051-5.
3. Karanastasi, A., Kazasis, F.G., Christodoulakis, S.: A natural language model for managing TV-anytime information in mobile environments. Personal and Ubiquitous Computing **9** (2005) 262–272 ISSN 1617-4917.
4. Soares, L.F.G.: MAESTRO: The declarative middleware proposal for the SBTVD. In: Proceedings of the European Interactive TV Conference. (2006) 538–541
5. Bulterman, D.C.: A rationale for creating a declarative interactive tv profile. In: Proceedings of the European Interactive TV Conference. (2006) 532–537
6. Jensen, J.F.: Interactive television: New genres, new format, new content. In: Second Australasian Conference on Interactive Entertainment. ACM International Conference Proceeding Series; Vol. 123, Sydney, Australia (2005) 89–96 ISBN 0-9751533-2-3.
7. de Resende Costa, R.M., Moreno, M.F., Rodrigues, R.F., Soares, L.F.G.: Live editing of hypermedia documents. In: Proceedings of the ACM Symposium on Document Engineering. (2006) 165–175
8. Chorianopoulos, K.: Virtual Television Channels: Conceptual Model, User Interface Design and Affective Usability Evaluation. PhD thesis, Athens University of Economic and Business (2004)
9. Cesar, P., Bulterman, D.C., Jansen, J.: Benefits of structured multimedia documents in idtv: The end-user enrichment system. In: Proceedings of the ACM Symposium on Document Engineering. (2006) 176–178
10. Olsen, D., Jefferies, J., Nielsen, T., Moyes, W., Fredrickson, P.: Cross-modal interaction using xweb. In: Proceedings of the ACM Annual Symposium on User Interface Software and Technology. (2000) 191–200
11. Nichols, J., Myers, B., Higgins, M., Hughes, J., Harris, T., Rosenfeld, R., Pignol, M.: Generating remote control interfaces for complex appliances. In: Proceedings of the ACM Annual Symposium on User Interface Software and Technology. (2002) 161–170
12. Beaudoin-Lafon, M.: Designing interaction, not interfaces. In: Proceedings of the International Working Conference on Advanced Visual Interfaces. (2004) 15–22
13. Bulterman, D.C., Jansen, J., Kleanthous, K., Blom, K., Benden, D.: Ambulant: A fast, multi-platform open source SMIL player. In: Proceedings of the 12th ACM International Conference on Multimedia, October 10-16, 2004, New York, NY, USA. (2004) 492–495 ISBN 1-58113-893-8.
14. Cesar, P., Bulterman, D.C., Jansen, J.: An architecture for end-user tv content enrichment. In: Proceedings of the European Interactive TV Conference. (2006) 39–47
15. ISO: Multi lingual information framework – multi lingual resource management. ISO/AWI 24616 (October 2006)
16. Cruz-Lara, S., Gupta, S., Gardia, J., Romary, L.: Multilingual information framework for handling textual data in digital media. In: Proceedings of the International Conference on Active Media Technology. (2005) 82–84