SMART DIFFUSION OF GOVERNMENT CONTENTS

Elena Sánchez-Nielsen[1], Francisco Chávez-Gutiérrez[2]
[1]*Dpto. E.I.O. y Computación, La Laguna University,*
38271 La Laguna, Spain

[2]*Parlamento de Canarias. C/Teobaldo Power, 7,*
38002 S/C de Tenerife, Spain

Keywords: E-Government services, Smart podcasting, thesauri.

Abstract: Podcasting provides a new way for government agencies of any size to extend their reach and information delivery to citizens. Currently, this technology is being used as an effective and inexpensive tool to deliver audio and video content on mobile devices and personal computers. Their essence is focused on creating multimedia content for audiences that want to listen/view when they want, where they want and how they want. In this paper, we focus our vision on to distribute public hearings and sessions of parliaments in an efficient and effective way to citizens. In order to provide customized content subscriptions according to citizens' preferences, we propose a semantic framework based on thesauri tools to describe the content of audiovisual information. In particular, we introduce three new components in traditional podcasting systems: (i) catalogue, (ii) fragmentation and (iii) customized delivery. Using these elements, we can establish how to present podcasting results according to customized user content features and how to reduce the size of multimedia segments to be downloaded and therefore, the quantity of information to be processed. This paper describes how this framework is being built on plenary sessions of the Parliament of the Canary Islands.

1 INTRODUCTION

Governments have made rapid progress worldwide in embracing ICT technologies for e-government in the past years. Examples of such improvements are: e-government portals of information and procedure services, points of access to EU legal documents, tax services and virtual offices of allegations. At the same time, the quantity and diversity of information available from public government sources is now quite large and increasing. Information available to citizens includes institutional communication (the official publication of laws and regulations), information on opportunities and promotional information. This significant amount of information is currently managed by traditional access paradigms that focus on retrieval of data using queries on structured database systems and information retrieval techniques. In order to enhance the access to information, we focus our vision on how to make this information more accessible to citizens. In particular, our research efforts are focused on to improve the efficiency and distribution of parliamentary sessions and public hearings to citizens, audience of business, nonprofits and institutions by means of podcasting technology and the Semantic Web paradigm.

Traditionally, Semantic Web paradigm has become popular in fields such as information integration, information retrieval on Internet and knowledge management. In this paper, we propose the use of Semantic Web paradigm and in more detail, the use of thesauri tools to achieve podcasting systems based on parliamentary contexts. In particular, we propose the using of resources based on thesauri, with the purpose of structuring, indexing and providing personalized audiovisual content. With this solution, we introduce a new approach to inform and distribute information in an efficient and effective way to citizens and other audiences that consists of four processes: (i) encode and archive, (ii) catalogue, (iii) fragmentation and (iv) customized delivery. Using these components, we can establish how to present podcasting results according to specific user content features and at the same time, we can establish how to reduce the size of multimedia segments to be downloaded and

Sánchez-Nielsen E. and Chávez-Gutiérrez F. (2006).
SMART DIFFUSION OF GOVERNMENT CONTENTS.
In *Proceedings of the International Conference on e-Business*, pages 294-299
DOI: 10.5220/0001424902940299

therefore, the quantity of information to be processed by end-users. With this approach, citizens, public authorities and other audiences can access more quickly and at lover cost to the information that they require. Moreover, this approach can be extended to other domains of the Public Administrations (PA's), whose goal is focused on the access to public information and reach of citizens and other authorities.

The remainder of this paper is structured in the following way. Section 2 gives an overview of podcasting technology and Semantic Web paradigm. Section 3 presents our content-awareness based podcasting framework and Section 4 provides concluding remarks and future work.

2 RELATED WORK

Podcasting is a recent and effective medium to deliver syndicated Web content (audio/video data) by content providers to consumer users. Interactions between providers and consumers consist of a two-step process. First, the content providers make accessible audio or video files on an available webserver, which are often referred to as one episode of a podcast. Then, the content provider acknowledges the existence of these files by referencing them in another file known as the feed. The feed is a machine-readable list of the episodes which may be accessed. This list is usually published in RSS format (RSS 2.0 Specification n.d.) (Really Simple Syndication), which provides other information, such as publish dates, titles, and accompanying text descriptions of the series and each of its episodes. The feed is typically limited to a short list of the most recent episodes. Second, a consumer user uses a software program called a podcatcher with the purpose of determining the location of the most recent episode and automatically downloads it to the user's computer or portable players. The downloaded episodes can then be played and replayed at anytime.

Currently, the using of a subscription feed of automatic delivery of new contents is what distinguishes a podcast from a simple download or real-time streaming. As a result, subscriptions to podcasting allow users to collect programs from a variety of sources for listening or viewing offline at anytime and anywhere. However, no personalized deliveries of specific contents have been included in traditional podcasting systems. In this context, Semantic Web (W3C: Semantic Web n.d.) paradigm becomes a key feature for ensuring an appropriate response to the user requests.

At present, the Semantic Web, the future of the current Web, is an extended web of machine-readable information and automated services that amplify the Web far beyond current capabilities. The explicit representation of the semantics underlying data, pages and other Web resources will enable a knowledge-based Web. The path to machine-processable data is to make the data smarter.

There are four different stages from data with minimal smarts to data embodied with enough semantic information for machines to make inferences about it:

- **Text and databases (pre-XML)**: this initial stage corresponds to applications to proprietary applications. Thus, the smarts are in the application and not in the data.
- **XML documents for a single domain**: data, in this stage, achieves application independence within a specific domain using structured categories defined to contain information about one aspect or attribute (e.g. publisher, subjects) of an information resource (e.g. book, document, image). An example of this context would be the personalized access to multi-version XML document repositories in an EGovernment scenario (Grandi et al, 2005, p. 281-290).
- **Taxonomies and thesauri based resources**: in this stage, data are composed from multiple domains and accurately classified in a hierarchically structured controlled vocabulary of terms that are used to describe information resources. Relationships and cross-references can be used to relate and thus combine data. Thus, data is now smart enough to easily discovered and sensibly combined with other data. Dynamic taxonomies (Sacco, 2000, p. 468-479) have been proposed as a tool to solve information access and dissemination needs of e-administrations (Sacco, 2005, p. 261-268). Intelligent thesauri such as WordNet (Fellbaum, 1998), an online lexical reference system, can also be used to support automatic text analysis and artificial intelligence applications.
- **Ontologies and rules**: in this stage, a structural specification is used to express complex relationships among concepts of a

domain of interest. New data can be inferred from existing data by logical rules. Different languages can be used to describe ontologies in a formal and explicit way such as RDF (W3C: Rdf. n.d.), DAML (The DARPA Agent Markup Language Homepage: Daml. n.d.), OIL (Ontoknowledge project: Oil n.d.), and OWL (W3C: Web ontology language n.d.).

Figure 1 shows the four different stages of the smart data continuum.

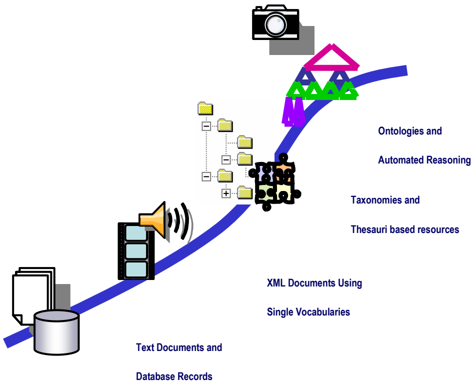


Figure 1: The smart data continuum.

3 SMART DIFFUSION OF AUDIOVISUAL CONTENTS

Our aim is to enable a customized podcasting video system to citizens in an e-government scenario related to parliamentary interventions, which involve different participants, debates, subjects to be discussed, and initiatives to be approved. Plenary sessions are the sessions to be included due to the public character such is specified at article 71 of the Parliament of Canary Islands Regulation (Reglamento del Parlamento de Canarias n.d.). With this aim, we present a framework that supports a new model of video podcasting to only deliver the fragments that end-users are interested. With this approach, we provide two main features to citizens, audience of business, nonprofits and institutions: (i) minimizing the time of searching a specific subject and (ii) reducing the size of data to be downloaded and processed.

In order to enable smart distribution of audiovisual contents of plenary sessions, we focus our vision on three goals: (i) the use of podcasting technology as mechanism of effective distribution of plenary sessions to citizens and other audiences, (ii) the use of Semantic Web paradigm with the purpose of providing semantic meaning to parliamentary videos and making available an appropriate response to the user requests and (iii) the automation of podcasting system by means of the use of a thesaurus that comprises all the activity areas of European Communities.

Figure 2 illustrates the resulting framework. This framework is composed by the following modules:

3.1 Encode and Archive

Two essential aspects need to be addressed by the encoding and archive module in order to provide high quality video to end-users: (i) clarity pictures and (ii) selection of adequate sound bitrates in order to produce sound quality near to FM radio. This situation involves high size of videos to be collected in the storage system.

The contents of plenary sessions are digitalized using Osprey video capture cards (Osprey video n.d.) in a Windows Media System, which is responsible of encoding and archiving the audiovisual content to a digital format for a posterior processing. Other systems equipped with audio and video capture cards are also used to record the different commissions' sessions that can take place at the same time.

Due to the increasing amount of multimedia material which is recorded from plenary and commissions' sessions, an efficient storage, management and administration of such material is of growing importance. This goal is achieved by using a storage area network (SAN). This storage medium is a dedicated network that is separate from LANs and WANs and it is used to interconnect the storage-related resources that are connected to one or more servers that contain the different applications related to podcasting system, e-government services and database systems.

The storage system consists of RAID storage systems as storage peripherals, fibre channel switches as interconnection equipment among all the components of the storage network, servers, backup devices to offload backup operations and interface cards as components to connect servers to storage network. Raid systems include data protection, fault tolerance and high scalability.

3.2 Catalogue

Catalogue process is achieved by thesauri experts who analyze the audiovisual content with the purpose of adding semantic meaning and different annotations. This process is carried out in real time

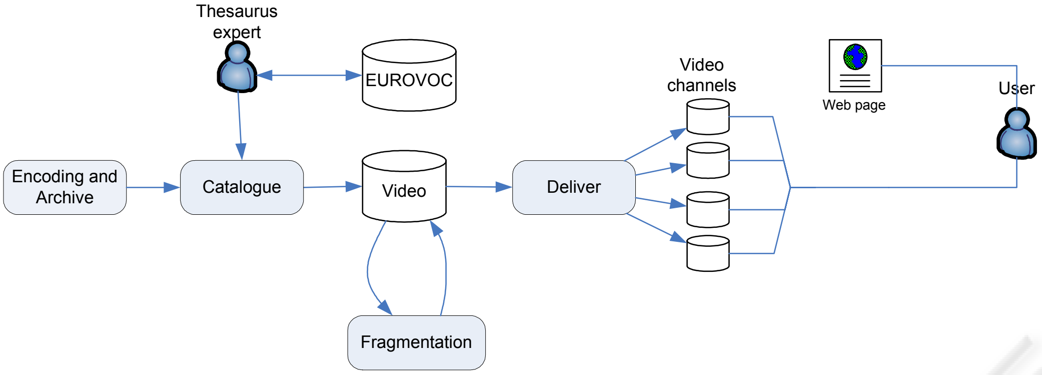


Figure 2: Smart diffusion of audiovisual contents.

while the sessions are taken place or more later when the sessions have concluded.

The annotations made by human thesauri expert identify the following aspects: (i) the name of the person who takes part, (ii) parliamentary initiative and (iii) the reason of the intervention. At the same time, the thesaurus expert describes the conceptual content that is taken place at every time by means of a hierarchy of concepts that are described by descriptors or terms. These descriptors are obtained from a thesaurus that is a basic model of ontology consisting of a database of hierarchical terms with semantic and generic relationships.

All the annotations are stored at a database system with other complementary information about the plenary session included by the system such as name of the session, date, and agenda and so on.

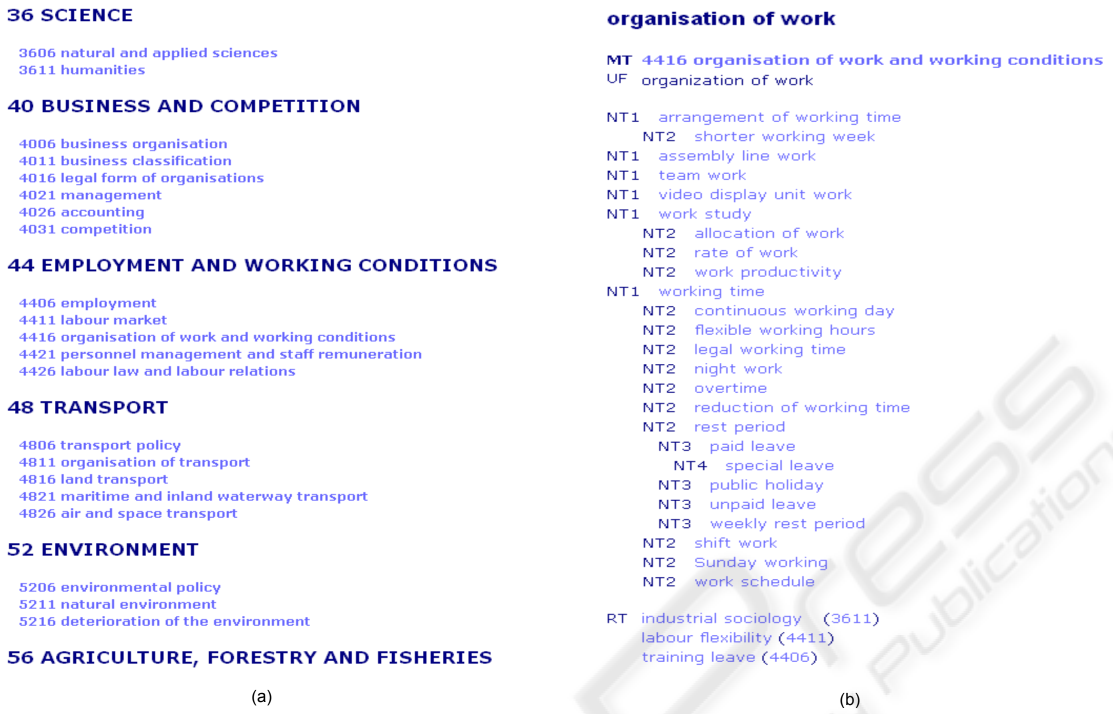
Multilingual Eurovoc thesaurus (The Office for Official Publications of the European Communities n.d.) is used by thesauri experts to catalogue in an ambiguous fashion the different pieces of the audiovisual content. Currently, Eurovoc thesaurus comprises:

- Descriptors: Those consist of words or expressions which denote in unambiguous fashion the constituent concepts of the domain covered by the thesaurus (e.g. implementation of the law).
- Non-descriptors: consist of words or expressions which in natural language denote the same concept as a descriptor (e.g. application of the law) or equivalent concepts (e.g. enforcement of the law, validity of the law).
- Semantic relationships: relationships based on meaning, firstly between descriptors and non-descriptors and secondly between descriptors.

In more detail, Eurovoc thesaurus comprises 21 fields, 127 microthesauri, 6439 descriptors, 6448 hierarchical relationship and 3501 associate relationship. The European Parliament, the Office for Official Publications of the European Communities, the national and regional parliaments in Europe, some national government departments and European organisations are currently using this thesaurus.

The 21 different fields covered by this thesaurus comprises all the areas of importance for the activities of the European institutions: politics, international relations, European Communities, law, economics, trade, finance, social questions, education and communications, science, business and competition, employment and working conditions, transport, environment, agriculture, forestry and fisheries, agri-foodstuffs, production, technology and research, energy, industry, geography and international organizations.

Every different subject of the audiovisual content is identified by a start and end tag. The thesaurus expert describes in unambiguous fashion the conceptual content between both tags using the descriptors from Eurovoc thesaurus that most clearly matches the conceptual concept. Each descriptor uses a two-tier hierarchical classification that comprises (i) a field, identified by two-digit numbers and titles in words, e.g.: 10 European Communities and (ii) a microthesauri, identified by four digit numbers, e.g.: 1011 Community Law, where the two first digits correspond to the field that contains the microthesaurus. Semantic relationships between descriptors comprise scope notes (some descriptors can be accompanied by notes, clarifying the meaning of the descriptor), microthesaurus relationships, equivalence relationships, hierarchical



36 SCIENCE

3606 natural and applied sciences
3611 humanities

40 BUSINESS AND COMPETITION

4006 business organisation
4011 business classification
4016 legal form of organisations
4021 management
4026 accounting
4031 competition

44 EMPLOYMENT AND WORKING CONDITIONS

4406 employment
4411 labour market
4416 organisation of work and working conditions
4421 personnel management and staff remuneration
4426 labour law and labour relations

48 TRANSPORT

4806 transport policy
4811 organisation of transport
4816 land transport
4821 maritime and inland waterway transport
4826 air and space transport

52 ENVIRONMENT

5206 environmental policy
5211 natural environment
5216 deterioration of the environment

56 AGRICULTURE, FORESTRY AND FISHERIES

(a)

organisation of work

MT 4416 organisation of work and working conditions
UF organization of work

NT1 arrangement of working time
    NT2 shorter working week
NT1 assembly line work
NT1 team work
NT1 video display unit work
NT1 work study
    NT2 allocation of work
    NT2 rate of work
    NT2 work productivity
NT1 working time
    NT2 continuous working day
    NT2 flexible working hours
    NT2 legal working time
    NT2 night work
    NT2 overtime
    NT2 reduction of working time
    NT2 rest period
        NT3 paid leave
            NT4 special leave
        NT3 public holiday
        NT3 unpaid leave
        NT3 weekly rest period
    NT2 shift work
    NT2 Sunday working
    NT2 work schedule

RT industrial sociology (3611)
labour flexibility (4411)
training leave (4406)

(b)

Figure 3: Eurovoc theasurus_ a) Some main fields of activity of European institutions and b) Microthesaurus associated to the field “Employment and Working Conditions”.

relationships and associate relationships.

Figure 3 illustrates graphically some fields of activity of Eurovoc thesaurus and in more detail the different descriptors associated to concept related to “organization of work and working conditions”.

3.3 Fragmentation

After the catalogue process has been performed, the annotations made by thesauri experts allow identifying the start and end of the different subjects discussed.

Each video of each plenary session is divided by an automatic process in smaller videos that will have the size equivalent to the subject discussed.

Once the fragmentation process has been finished, each video fragment is assigned to the corresponding channel. Each fragment of video corresponds to one or several fields of activity of Eurovoc thesaurus. These fields of activity are computed by means of recovering the descriptors assigned by thesauri experts at catalogue process and by using hierarchical structure of the thesaurus Eurovoc with the purpose of identifying the root layer where this descriptor is associated.

3.4 Customized Delivery

The 21 different fields of Eurovoc thesaurus define the available channels that can be syndicated by citizens and other audiences.

In order to citizens know the available channels; a list with the different channels will be published in a web page. Each channel represents a feed, that is, a XML file based on RSS specification (RSS 2.0 Specification n.d.). In particular, this specification is the format used in podcasting that allows transforming the way millions of people consumes news and information. The growth adoption rate of RSS allows people rarely read information directly from a website.

In our podcasting architecture, each day, the system will be responsible of generating the RSS file of each channel in order to reflect the availability of the new episodes. The 10 last episodes of each channel will be included. In this context, users may select the appropriated RSS files corresponding to the channels they are interested.

Once syndication to a channel has been achieved, the software program (podcatcher) of every citizen or institution subscribed computes the localization of the most recent episodes of video for each user subscribed and automatically will download it to the user’s computer or portable player. The downloaded episodes can then be played, replayed, or archived as any other computer file.

4 CONCLUSIONS AND FUTURE WORK

In this paper, we present a new electronic service delivery to citizens, audience of business, nonprofits and institutions. We focus our vision on how to make information based on audiovisual contents more accessible to several audiences. In particular, our research efforts are focused on to improve the efficiency and distribution of public hearings and parliamentary sessions by means of podcasting technology and the Semantic Web paradigm. Podcasting technology is used as a mechanism to distribute audiovisual content, which can be played and replayed offline at anytime. Semantic Web paradigm gives a promising solution to customized delivery. The explicit representation of the semantics underlying podcasting system enables a qualitatively new level of service and provides a new approach to create smart podcasting to deliver customized information to different audiences and automate the assignation of information to different channels in the podcasting system.

Currently, we have developed encode and archive and catalogue process.

Future work will be focused on developing the remainder processes and extending the framework proposed in order to incorporate query formulation by end-users who want to search the spoken words related to main topics within any audio or video file. Also, new optimization techniques that produce video content according to different audience players will be evaluated.

REFERENCES

Grandi, F, Mandreoli F, Martoglia R, Ronchetti E, Scalas MR and Tiberio P, 2005. *Personalized access to multi-version Norm Texts in an eGovernment scenario*, M.A. Wimmer et al. (EDs.): EGOV 2005, LNCS 3591.

Sacco, G.M, 2000. *Dynamic Taxonomies: A Model for Large Information Bases*. IEEE Transactions on Knowledge and Data Engineering 12, 2.

Sacco G.M, 2005. *Guided Interactive Information Access for E-Citizens*. M.A. Wimmer et al. (EDs.): EGOV 2005, LNCS 3591.

Fellbaum, C. (ed.), 1998. *WordNet - An Electronic Lexical Database*. MIT Press.

W3C: Rdf. Retrieved from http://www.w3.org/RDF/

The DARPA Agent Markup Language Homepage: Daml. Retrieved from http://www.daml.org/

Ontoknowledge project: Oil. Retrieved from http://www.ontoknowledge.org/oil/

W3C: Web ontology language. Retrieved from http://www.w3.org/2004/OWL/

RSS 2.0 Specification. Retrieved from http://blogs.law.harvard.edu/tech/rss

The Office for Official Publications of the European Communities. Eurovoc Thesaurus. Retrieved from http://europa.eu.int/celex/eurovoc/

W3C: Semantic Web. Retrieved from http://www.w3.org/2001/sw/

Reglamento del Parlamento de Canarias. Retrieved from http://www.parcan.es/pub/reglamento.pdf

Osprey video. Retrieved from http://www.viewcast.com/products/osprey.html