CAEPIA 2011

Tecnologías de Linked Data y sus aplicaciones en España

Improving Television Program Guides by Accessing Linked Data Resources in OntoTV System

José Luis Redondo-García, Álvaro E. Prieto, Adolfo Lozano-Tello
Universidad de Extremadura. Escuela Politécnica, Cáceres, Spain
{jluisred, aeprieto, alozano} @unex.es

Software Engineering Group

QUERCUS

Índice

Introducción

Motivación:

- Existen cada vez un número mayor de **plataformas** y canales de televisión: TDT, Cable, Internet
- La información sobre los contenidos emitidos que los proveedores suministran es **escasa**

- El televidente necesita más información para **decidir** qué ver dentro de la amplia oferta televisiva disponible

Introducción

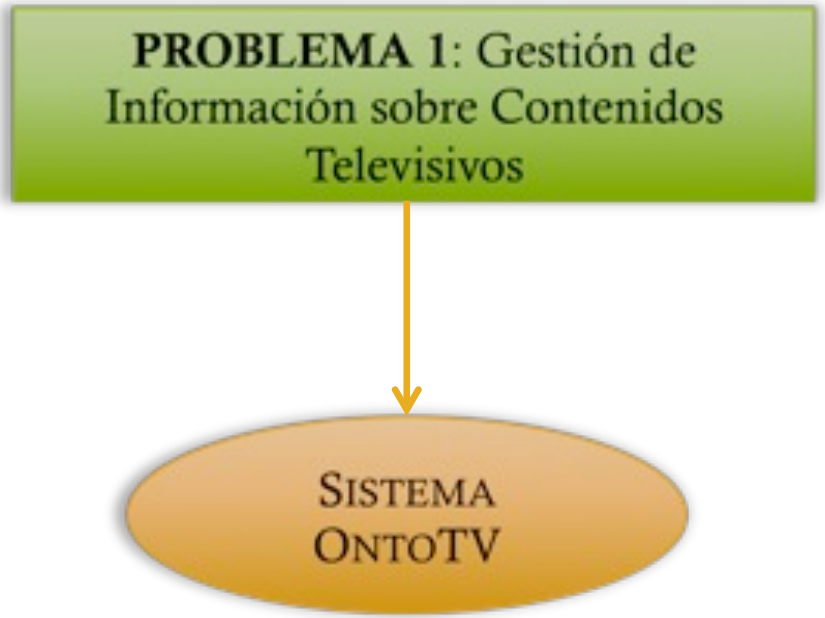
Motivación:

- Ciertas funcionalidades como las **búsquedas** personalizadas o las **recomendaciones**, resultan cada día más familiares y cercanas para el usuario en Internet

- Existen sistemas de televisión digital con amplias posibilidades en programación de aplicaciones e interactividad, como **MHP** o **GoogleTV**

Introducción

Motivación:



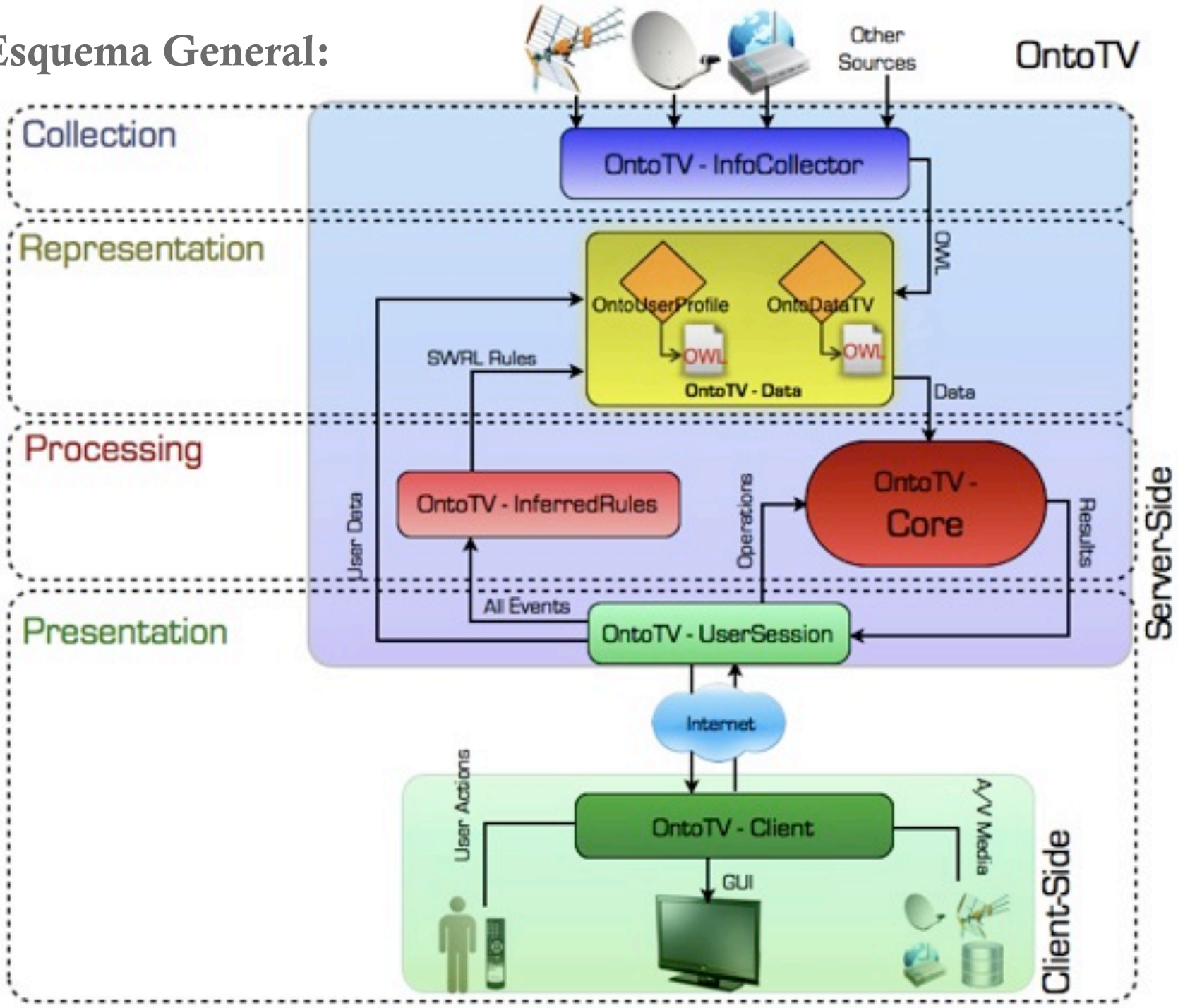
El Sistema OntoTV

Sistema OntoTV

Principales Características:

- Utiliza un **modelo ontológico** para la representación del conocimiento disponible sobre programas de televisión. Esto mejora las operaciones de búsqueda y recomendación
- Tiene un diseño de tipo **modular**.
- Este diseño permite que el sistema sea **flexible** ante futuros cambios
- Implementa una arquitectura de tipo **cliente-servidor**. (Como la mayor parte de servicios interactivos en la actualidad)
- Es **independiente** del sistema de televisión utilizado (MHP, GoogleTV, etc). Ofrecer un servicio televisivo **global**.

Esquema General:

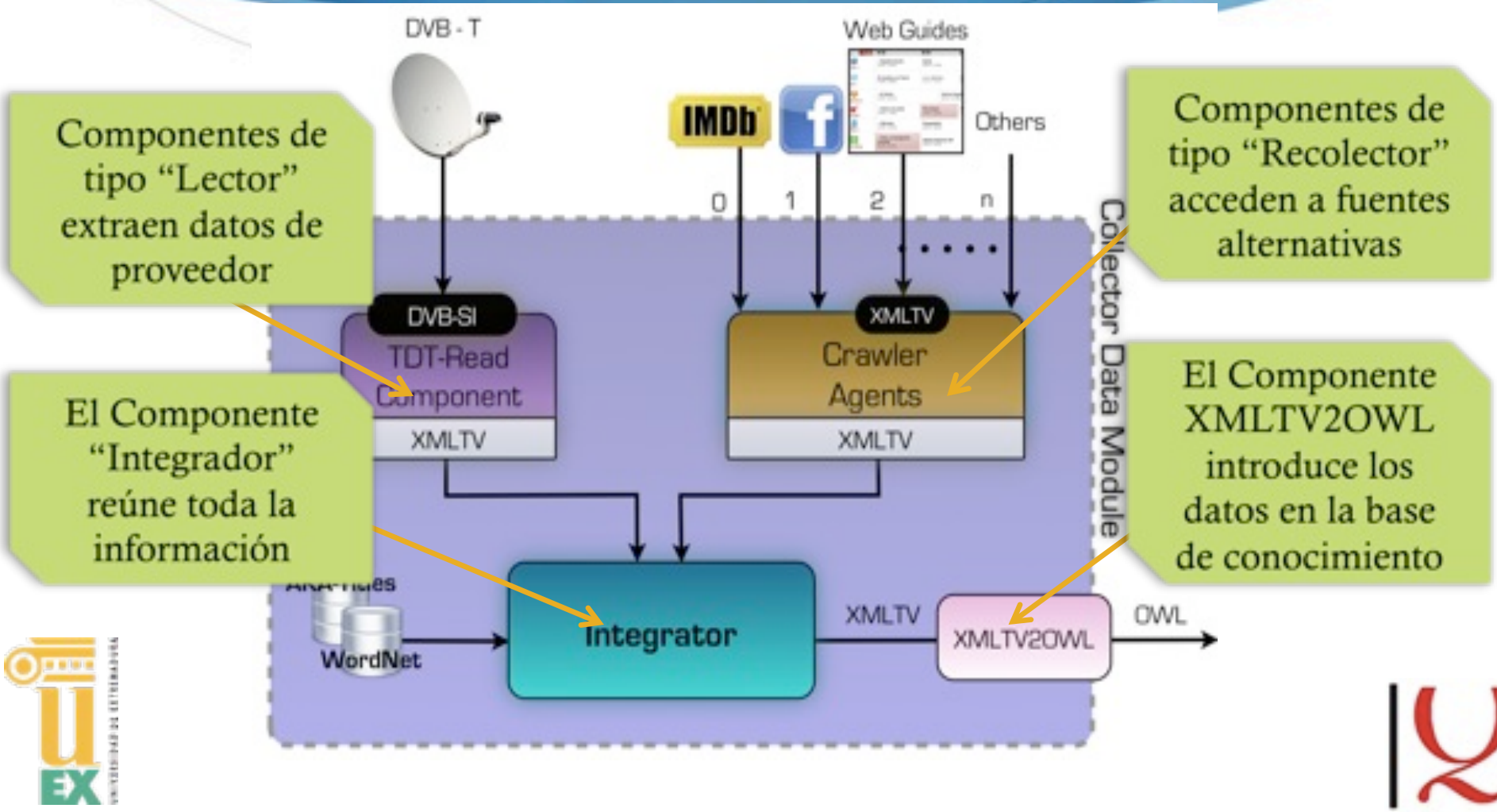


OntoTV: Recolección

Formato de los Datos de Entrada:

- TV-Anytime
  - Específico para televisión
  - Escaso soporte y desarrollo en la actualidad
- MPEG7
  - Extenso en su especificación
  - Propósito demasiado general para TV
- XMLTV
  - Muy simple. No describe preferencias de usuario.
  - Suficiente, dada la cantidad de información sobre contenidos actual

OntoTV: Recolección

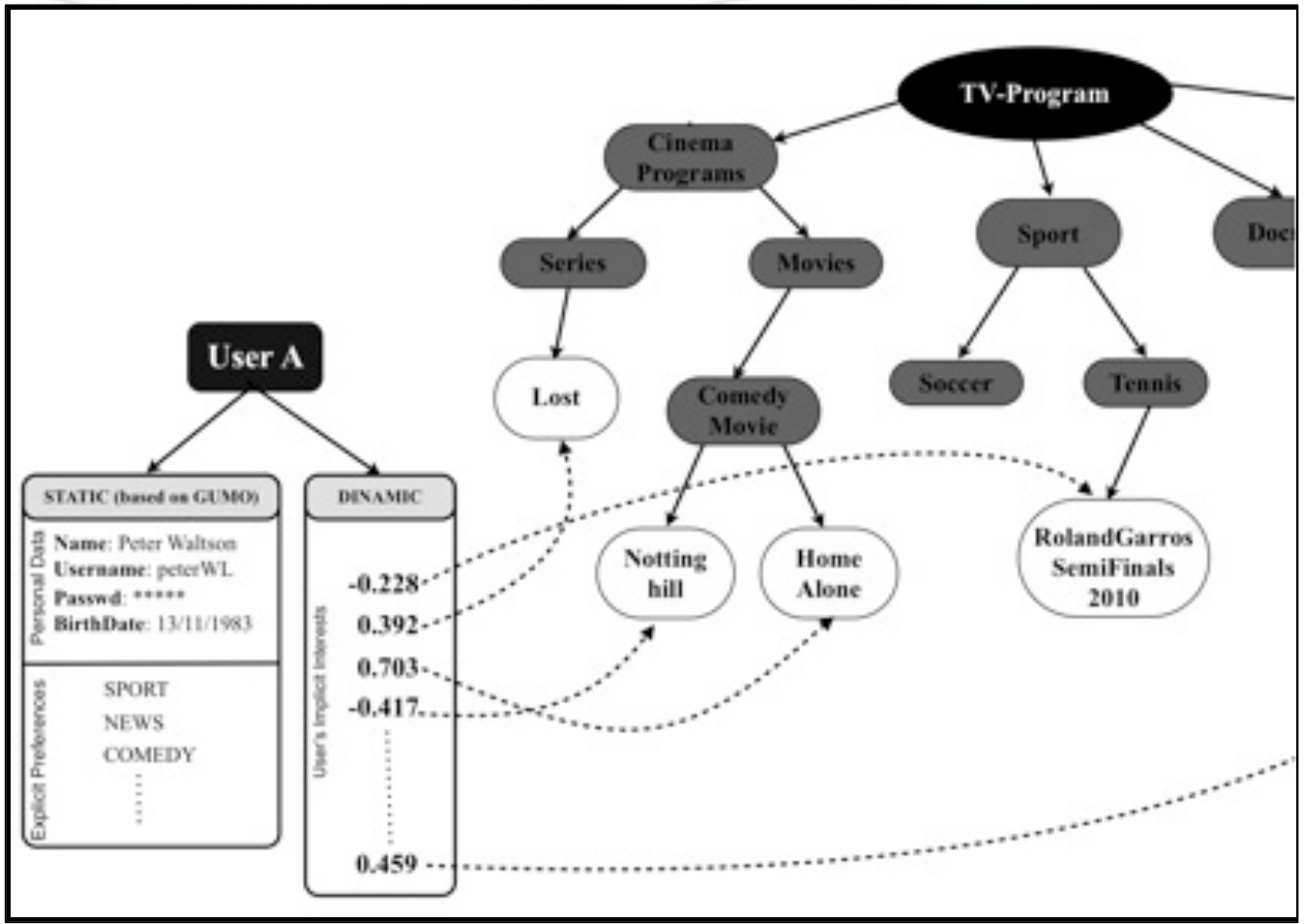


OntoTV: Representación

Representación de la información: Ontologías

- Ontología de datos televisivos, Onto-TVData
  - Basada en la propuesta por AVATAR
    - http://avatar.det.uvigo.es/software.html.
  - Jerarquía de categorías (deportes, películas, series…).
- Ontología de televidentes, Onto-UserProfile
  - Parte estática: Basada en GUMO (General User Model Ontology) Preferencias de Usuario Explícitas
  - Parte dinámica: Preferencias Implícitas.
    - Pares [Referencia, Valor]
    - Aceptación, porcentaje consumido, afinidad con preferencias explícitas, similitud jerárquica.
  - Reglas de producción para expresar hábitos. Motor de Inferencia.

Representación de Datos



OntoTV: Procesamientos

Procesamiento (En fase de desarrollo)

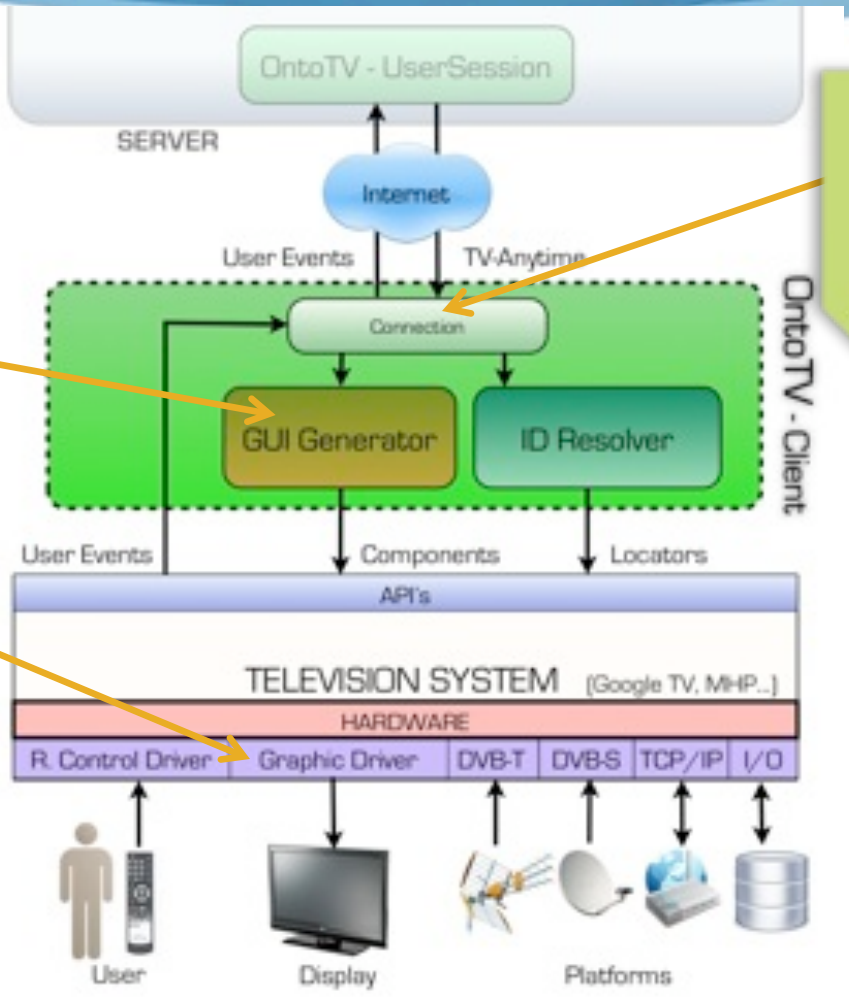
- Inferencia de hábitos de usuario (OntoTV-InferredRules)
  - Técnicas de Data Mining y Algoritmos de Aprendizaje
- Operaciones avanzadas para el Televidente:
  - Funcionalidades de Búsqueda
    - SPARQL
  - Funcionalidades de Recomendación
    - Recomendador Basado en Contenido-Colaborativo
    - Recomendador Implícito-Explícito
  - Otras: Generación de Guías Personalizadas

OntoTV: Presentación

"GUI Generator" crea las Interfaces de Usuario

El Sistema de Televisión Digital ejecuta el Cliente y le proporciona API's

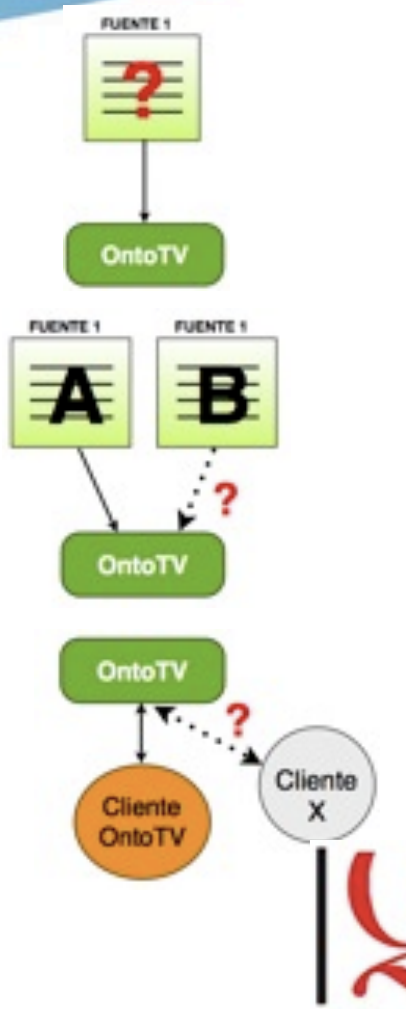
"Connection " establece la conexión TCP/IP con servidor.



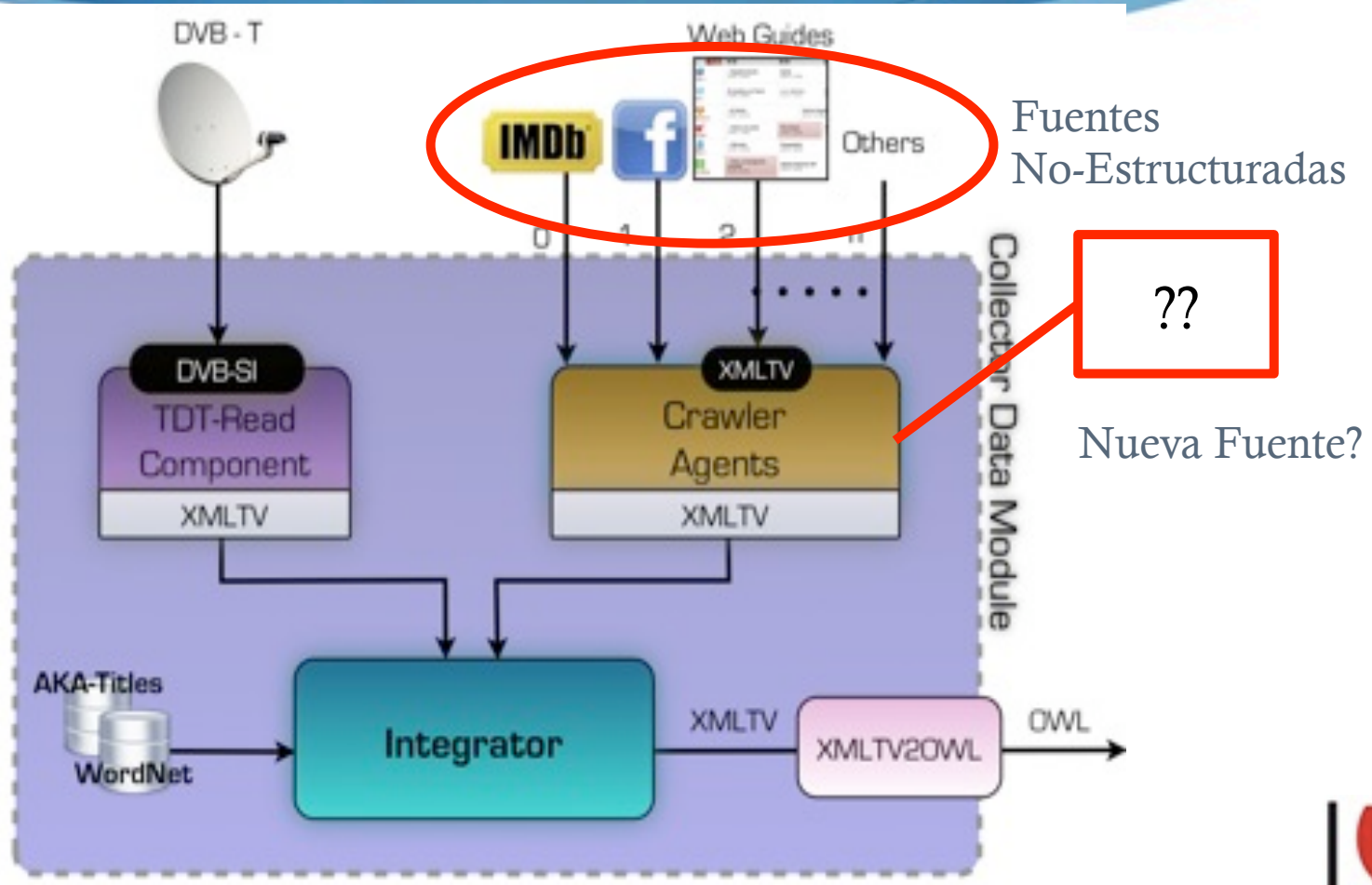
Problemas en OntoTV

Problemas:

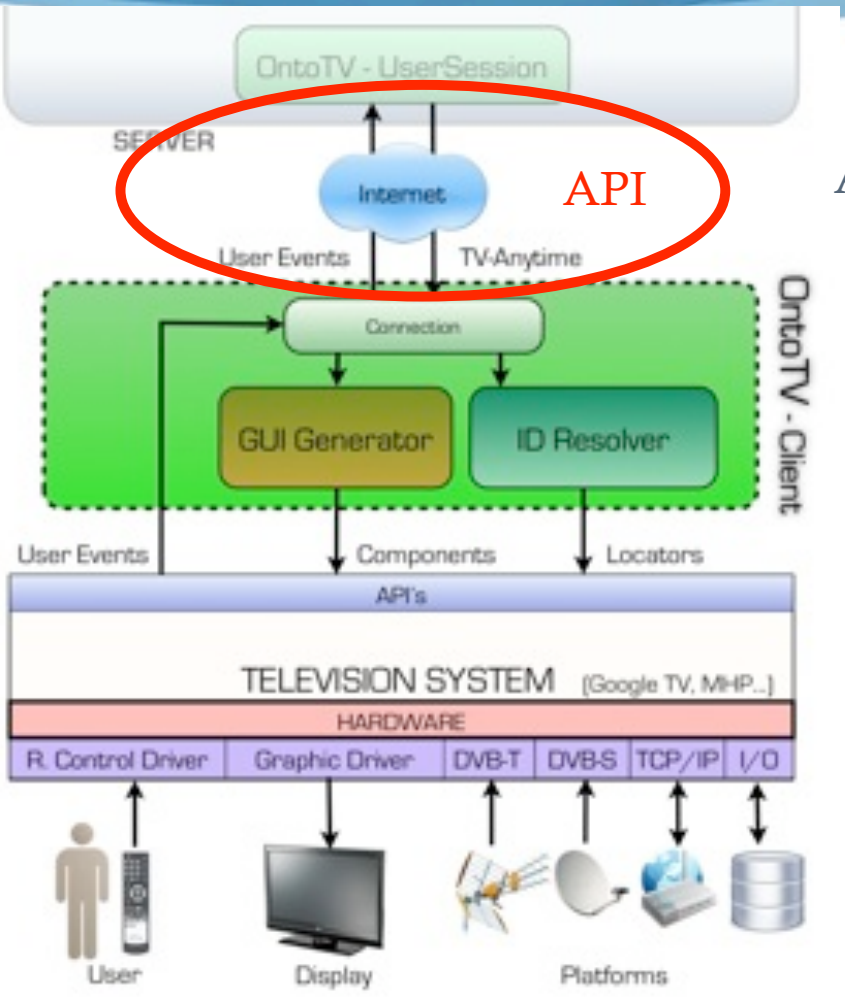
- En las fuentes que sí ofrecen información, esta no esta convenientemente **estructurada** ➔ Difícil de interpretar
- Cada fuente de información ofrece distintos métodos de acceso a sus datos ➔ Dificultades al **integrar** nuevas fuentes
- OntoTV sólo permite que clientes **compatibles** con él accedan a la información. ➔ Dificultad para reutilizar datos



Problemas en OntoTV



Problemas en OntoTV



A implementar por todos los clientes

Consumo y Publicación de Datos

Problemas:

Linked Data en OntoTV

Aplicar Linked Data en OntoTV

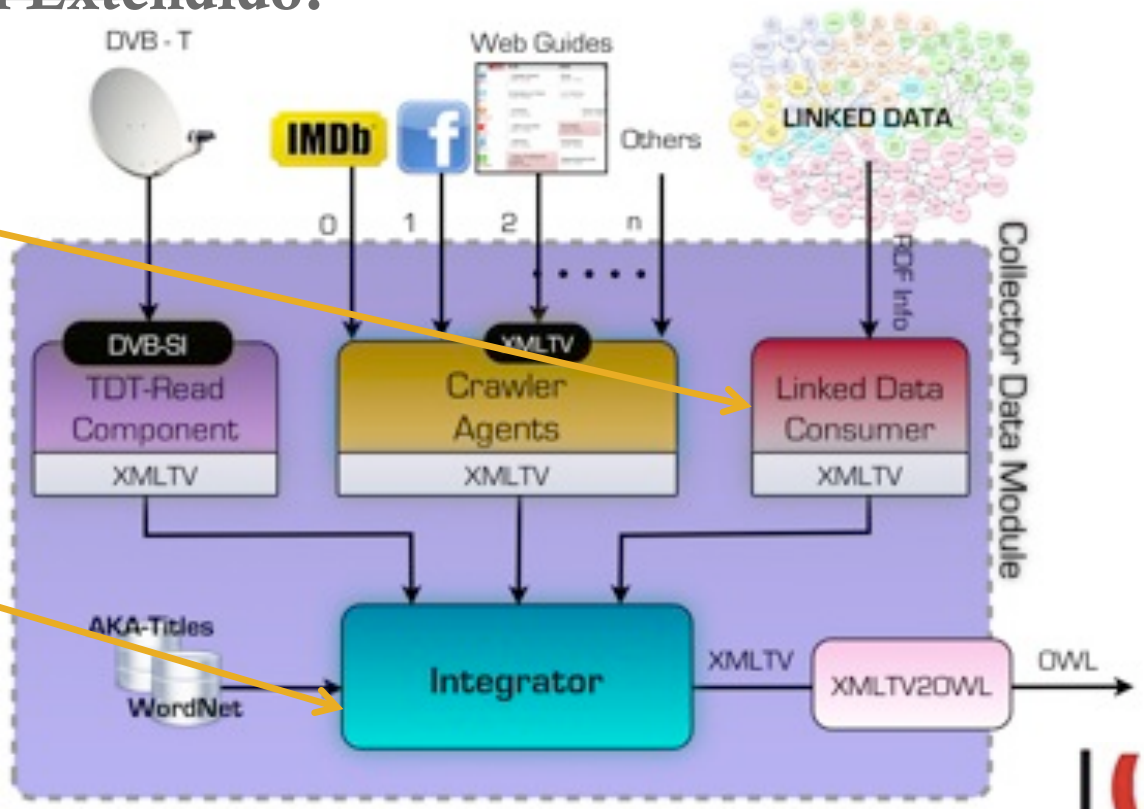
- Identificar **fuentes** Linked Data sobre dominio televisivo
- Modificar la **arquitectura** del sistema OntoTV para soportar consumo y publicación de información en la Web de Datos.
  - Implementar un nuevo componente para el **consumo** de información Linked Data sobre películas.
  - Diseñar un mecanismo para **publicar** información sobre contenidos de OntoTV en Linked Data.

Aplicando LD en OntoTV

Módulo de Recolección Extendido:

El componente "Linked Data Consumer" accede a información extra en la Web de Datos (películas)

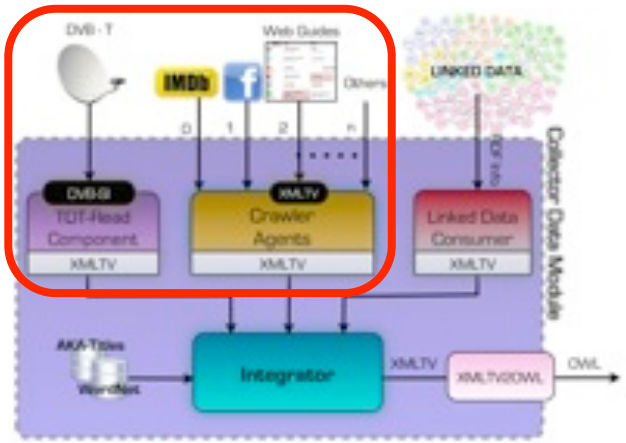
"Integrator" identifica descripciones de un tipo particular de contenidos (películas) y solicita al información extra.



LD en OntoTV:
Proceso de Recolección

FASE 1: Captura de Datos

- Se leen guías de programación oficiales de plataformas Digitales disponibles. Ejemplo: guía de TDT enviada sobre estándar DVB-T
- Se accede a guías de programación alternativas, normalmente disponibles en Internet.
  - La página web de "La Guía TV" xmltv-0.5.59
  - La página web de "Mi Guia TV" xmltv-0.5.59
  - Guía de Programas de Windows Media Center



LD en OntoTV:
Proceso de Recolección

Fase 1: Captura de Datos

```
<!-- DTT READER -->
<programme channel="15" start="20101214210940" stop="2
    <title>Blade Runner</title>
    <sub-title></sub-title>
```

```
<!-- LAGUIATV.COM -->
<programme start="20101214220000 +0100" channel="Cl
    <title lang="es">El cine de La 2: Blad
    <category lang="es">pelicula</category
</programme>
```

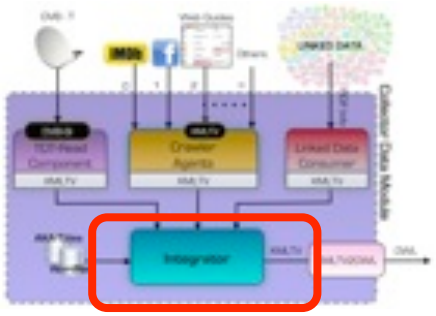
```
<!-- MIGUIATV.COM -->
VED -->
```

```
<!--WINDOWS MEDIA CENTER -->
<programme start="20101214220000 +0100" stop="20101214
    <title lang="es">El cine de La 2</title>
    <desc lang="es">Espacio que incluye la emisión de
    <date>20070427</date>
    <category lang="es">Otro</category>
    <category lang="es">Película</category>
    <length units="minutes">120</length>
</programme>
```

LD en OntoTV:
Proceso de Recolección

FASE 2: Fusión

- Identificación de descripciones pertenecientes a un mismo contenido (Agrupación de instancias):
  - **Similitud espacio-temporal del contenido (S_{E-T}).** Si un programa pertenece al mismo canal, y comienza y finaliza a una hora similar, alta similitud.
  - **Similitud en los títulos (S_T).** Funciones de comparación relativa de cadenas (Damerau–Levenshtein), y títulos alternativos de IMDB. Wordnet.
  - **Similitud global (S_{global})**. Palabras que aparecen en ambas a la vez, sin importar el lugar exacto en que lo hagan



$$C_A \approx C_B \Leftrightarrow P_1 \cdot S_{E-T} + P_2 \cdot S_{Titulo} + P_3 \cdot S_{Global} \geq Umbral$$

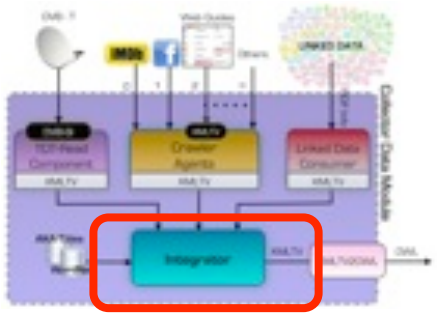
$$S_x \in [0,\ 1],\quad P_1 + P_2 + \ldots + P_n = 1,\quad 0 < Umbral \leq 1$$

P_1=0.5, P_2=0.35, P_3=0.15, Umbral = 0.65

LD en OntoTV:
Proceso de Recolección

FASE 2: Fusión

- Unificar las descripciones pertenecientes a un mismo grupo.
  - **El campo sólo aparece en una de las descripciones, se toma inmediatamente.**
  - **El campo aparece en varias descripciones:**
    - Si es un atributo múltiple, se añaden todas sus apariciones.
    - Si sólo debe aparecer una vez en cada descripción:
      - Si es posible, se concatenan las distintas versiones
      - Si no lo es, se elige fuente de mayor prioridad.



  Al final se obtiene una descripción única de cada contenido

- Se elabora una Guía Única. Todas las descripciones generadas se almacenan en un nuevo archivo XMLTV.

LD en OntoTV:

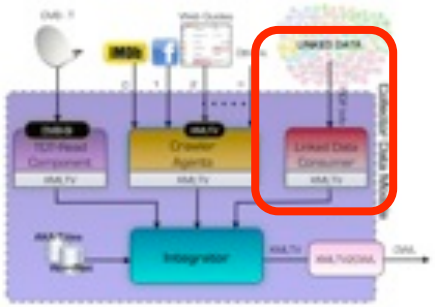
Nuevas Fases del Proceso de Recolección

LD en OntoTV:

Nuevas Fases del Proceso de Recolección

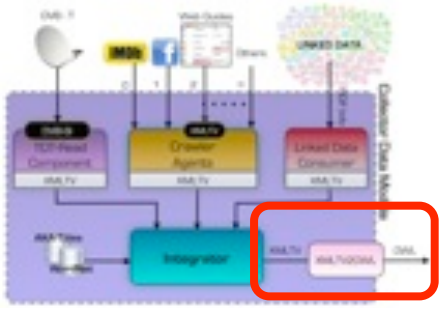
FASE 3: Enriquecer Descripciones de Películas Accediendo a LD

- Si el campo "category" es de tipo película:
  - Se solicita a "Linked Data Movies" más información sobre la misma.
  - Se rellenan los campos oportunos con la información conseguida.



FASE 4: Almacenar la Información en la Base de Conocimiento:

- Es llevada a cabo por el componente XMLTV2OWL
- Convierte elementos XMLTV de tipo "programme" en Instancias de la Ontología de Contenidos.



LD en OntoTV:

Acceso a Información Linked Data sobre Películas

Estrategias para Consumo de Información sobre Películas:

- Realizar operaciones de búsqueda en el dataset LinkedMDB utilizando **SPARQL**.
- Utilizar el patrón de consumo "**Crawling**": Jena TDB + LDSpider
  - Propuesto por Tom Heath and Christian Bizer en "Linked Data: Evolving the Web into a Global Data Space"
- Acceso al Integrador Semántico **SIG.MA**:
  - Recoge información de multitud de fuentes LD.
  - Coste computacional bajo.
  - Información con índice de actualización adecuado.

LD en OntoTV:

Acceso a Información Linked Data sobre Películas

Paso 1: Conseguir descripciones en formato RDF:

- Resolver **URIs HTTP** y recolectar así el RDF asociado.
- En el caso de **SIG.MA**, http://sig.ma/search?q=moviename

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF
    xmlns:foaf="http://xmlns.com/foaf/0.1/"
    xmlns:dc="http://purl.org/dc/terms/"
    xmlns:owl="http://www.w3.org/2002/07/owl#"
    xmlns:sigma="http://sig.ma/property/"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
<rdf:Description rdf:about="http://sig.ma/sigma/search?q=Blade+Runner">
    <dc:title>Description of Blade Runner | SIGMA</dc:title>
    <rdfs:label>Blade Runner</rdfs:label>
    <rdfs:comment>Blade Runner is a 1982 American sci-fi film, directed by Ridley S
    <foaf:depiction>http://upload.wikimedia.org/wikipedia/commons/5/53/Blade_Runner.
    <foaf:primaryTopic>
        <rdf:Description rdf:about="http://sig.ma/sigma/search?q=Blade+Runner#this":
            <sigma:otheruses4Property>
                <rdf:Description rdf:about="http://dbpedia.org/resource/William_S._
                    <rdfs:label>William S. Burroughs</rdfs:label>
                    <rdfs:seeAlso rdf:resource="http://dbpedia.org/resource/Blade_Ru
                </rdf:Description>
            </sigma:otheruses4Property>
            <sigma:otheruses4Property>Blade Runner (film)</sigma:otheruses4Property>
            <sigma:otheruses4Property>the unrelated film</sigma:otheruses4Property>
```

LD en OntoTV:
Acceso a Información Linked Data sobre Películas

Paso 2: Extraer información de los Ficheros RDF utilizando SPARQL:

- Se utiliza la librería **Jena ARQ** para ejecutar consultas en SPARQL sobre el fichero conseguido en la fase anterior.

```
PREFIX sigma: http://sig.ma/property/
PREFIX rdfs: http://www.w3.org/2000/01/rdf-schema#
SELECT ?director ?name
WHERE {
    ?film  sigma:director ?director.
    ?director rdfs:label ?name.
Retrieving info from SIG.MA...
-----------------------------------------------------------------------
| director                                    | name                  |
=======================================================================
| <http://dbpedia.org/resource/Ridley_Scott> | "Sir Ridley Scott" |
-----------------------------------------------------------------------
```

LD en OntoTV:
Acceso a Información Linked Data sobre Películas

Paso 3: Navegando hacia otros Datasets:

- Si la información en SIG.MA es **insuficiente** ➔Acceder a documentos alternativos siguiendo enlaces RDF

- Para rescatar más información sobre **Director** ➔ Aplicar de nuevo Paso 1 y 2 sobre la URI asociada al mismo (http://dbpedia.org/resource/Ridley_Scott)

```
PREFIX dbpedia-owl: <http://dbpedia.org/ontology/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT ?date
WHERE {
    ?director  dbpedia-owl:birthDate ?date.
    }
```

```
Finding for more information in http://dbpedia.org/resource/Ridley_Scott
-------------------------------------------------------------
| date                                                      |
=============================================================
| "1937-11-30"^^<http://www.w3.org/2001/XMLSchema#date> |
-------------------------------------------------------------
```

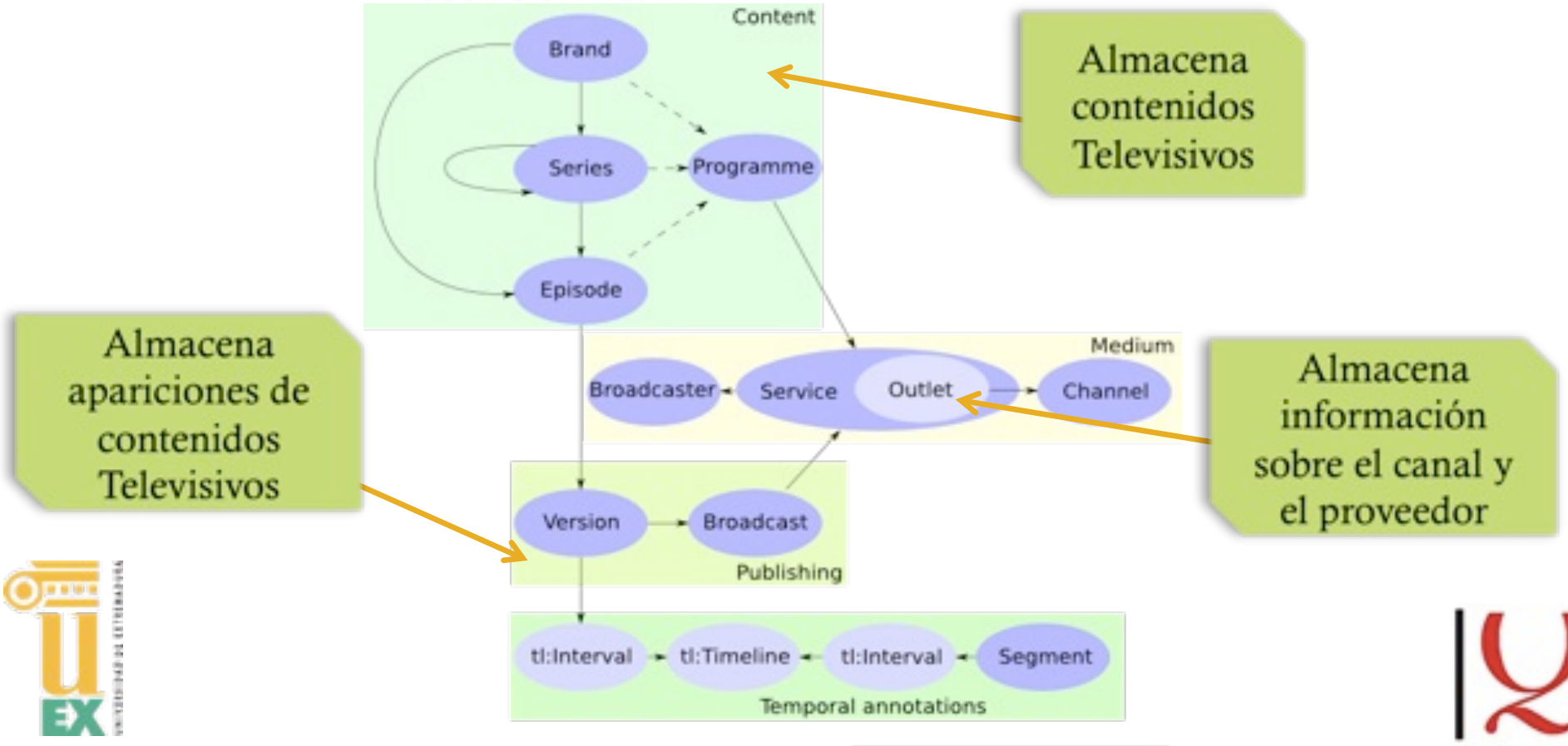
LD en OntoTV:
Acceso a Información Linked Data sobre Películas

Otros Posibles Ítems para películas:

Item	Elemento XMLTV	Propiedad en SIG.MA
Language	tv.programme.language	<sigma:language>
Length	tv.programme.length	<sigma:runtime>
Country	tv.programme.country	<sigma:country
Rating	tv.programme.rating	<sigma:ratings>
Director	tv.programme.credits.director	<sigma:director>
Actor	tv.programme.credits.actor	<sigma:starring>
Writer	tv.programme.credits.writer	<sigma:writer>
Producer	tv.programme.credits.producer	<sigma:producer>
Composer	tv.programme.credits.composer	<sigma:music_composer>
Image	tv.programme.icon	<sigma:picture>

LD en OntoTV:
Publicación de Datos Televisivos

Ontología de Dominio para Televisión: Ontología "BBC Programmes"



LD en OntoTV:
Publicación de Datos Televisivos

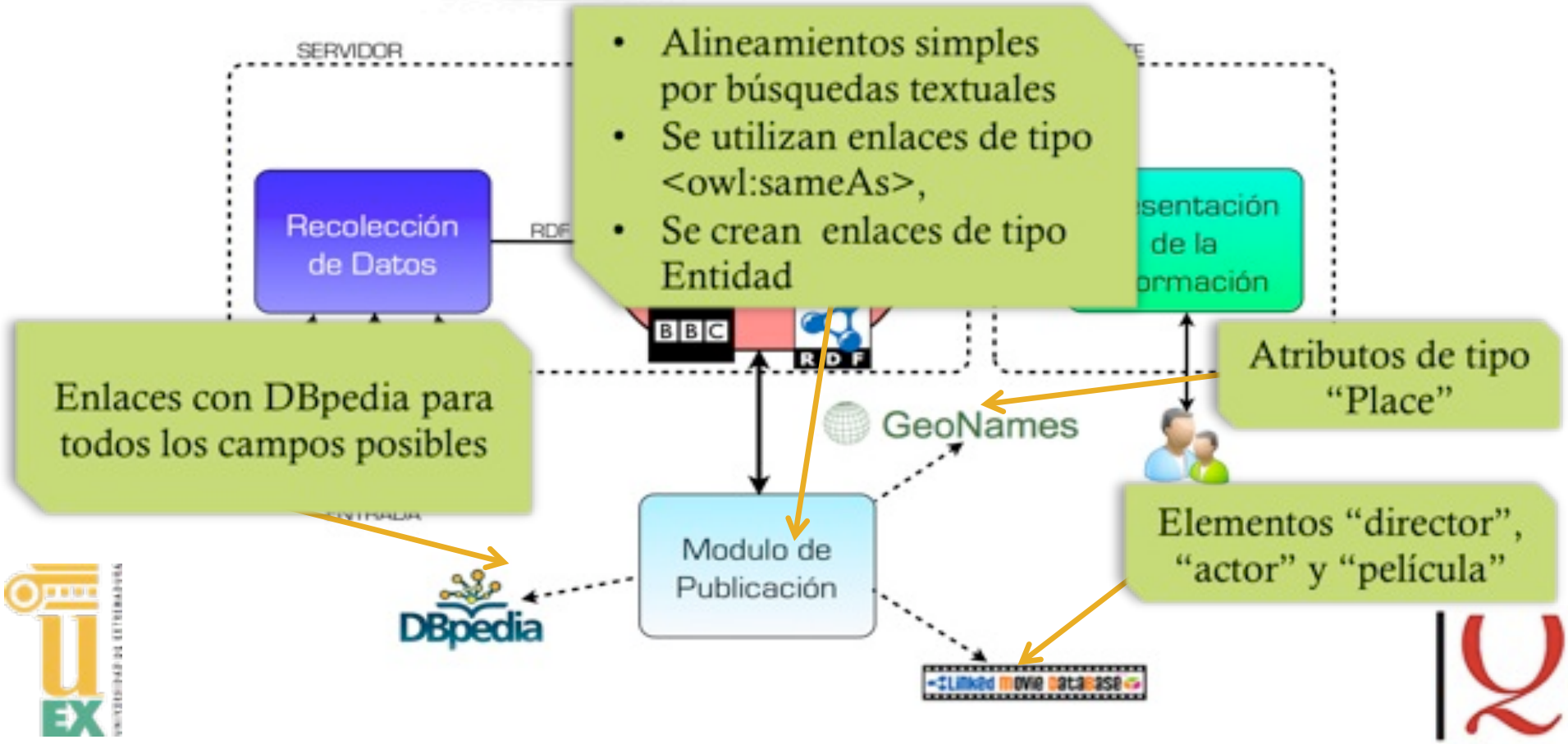
Inserción de datos en la ontología. Generación de código RDF:

- El sistema utiliza Ontologías ➔ Mayor sencillez al publicar en LD
- En OntoTV, el componente **XMLTV2OWL** realiza la conversión XMLTV a instancias de BBC Programmes
- Para este caso:
  1. Los elementos "<channel>" en XMLTV ➔ instancias de la clase Channel de la ontología de la BBC
  2. Para elementos de tipo "programme" en XMLTV:
     a. Se crea una instancia de tipo Programme en la ontología de datos televisivos del sistema (Si previamente no existía).
     b. Se crea una instancia de la clase Versión y se la asocia la instancia de programas seleccionada en el punto 2.a. (Se elimina la más antigua)

Aplicando LD en OntoTV:

Publicación de Datos Televisivos

Nuevo Módulo para la Publicación de Datos Televisivos en Linked Data:



Resultados, Aportaciones y Conclusiones

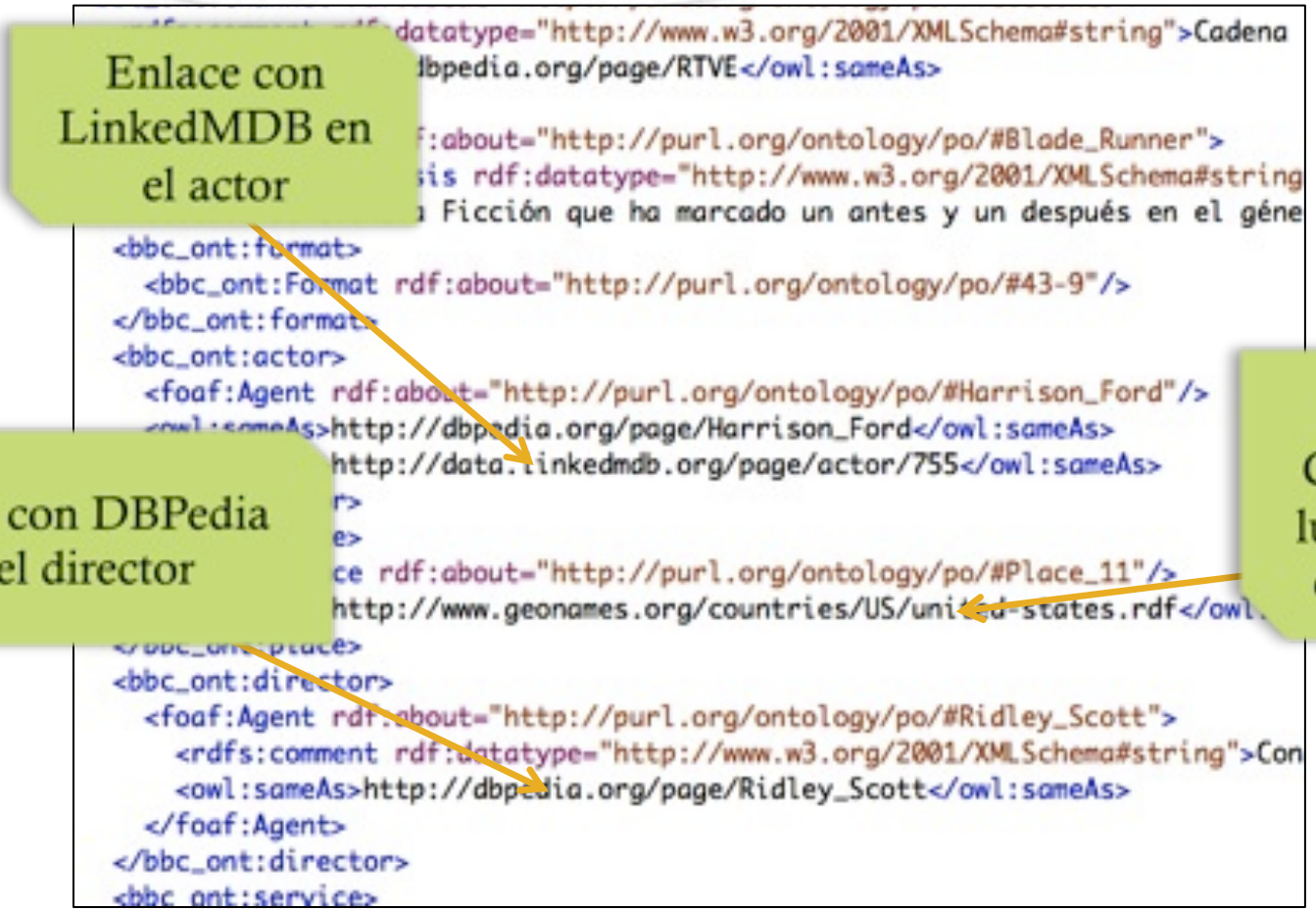
Resultados:

Datos Recolectados sobre Contenidos de Televisión


```
<!-- FINAL XMLTV -->
<programme start="20101214220000 +0100" stop="20101214235000 +0100" channel="fu
    <title lang="es">El
    </credits>
    <icon src="http://getmovielink.com/images/covers/BladeRunner.jpg" />
    <length units="minutes">120</length>
</programme>
```

Resultados:

Instancias en RDF Publicadas



Conclusiones:

- Actualmente no se dispone de sistema que **gestione** adecuadamente la información sobre contenidos televisivos.

- El sistema **OntoTV** ha sido diseñado para dar solución a este inconveniente. Sin embargo:
  - Fuentes de naturaleza No Estructurada, proceso de recolección computacionalmente costoso.
  - Sólo clientes de OntoTV pueden acceder a la información recolectada.

Conclusiones:

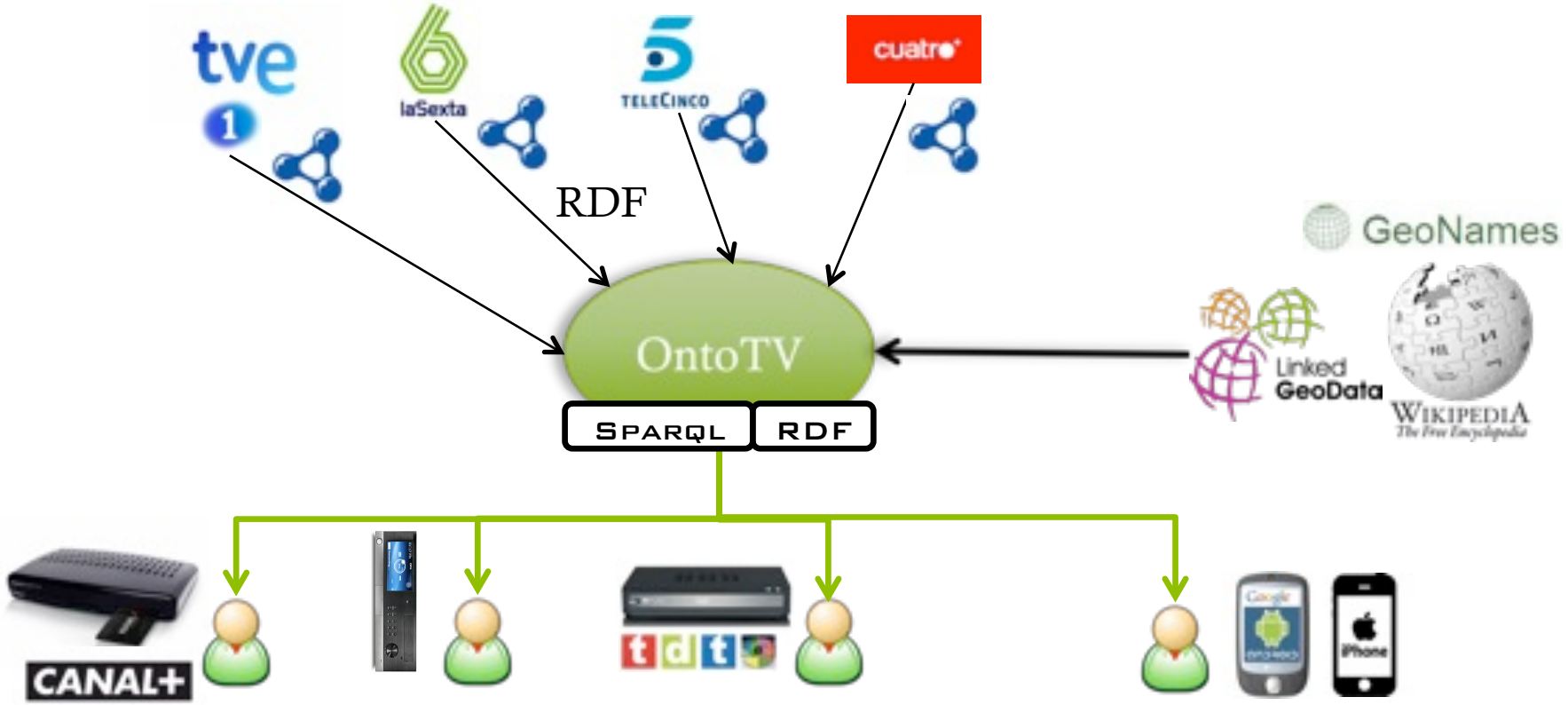
- Para solucionar estos inconvenientes se han aplicado **principios** de Linked Data al sistema OntoTV :
  - Se ha **extendido** la arquitectura del sistema para soportar metodologías Linked Data, manteniendo la lógica de funcionamiento del resto de OntoTV.
  - Se ha codificado el componente "**LinkedData Movies**" para información adicional sobre películas en la Web de Datos ➔ **Aplicable** a otros tipos de contenidos.
  - **Mecanismo** válido para publicar contenidos de televisión en la Web de Datos utilizando la ontología "BBC Programmes".
- Aplicación de **tecnologías semánticas** ha sido beneficiosa para consumo y publicación.
- Seguir **mejorando**:
  - Transformación de datos XMLTV a instancias de la ontología.
  - Procesos de alineamiento de instancias de Programas

Líneas Futuras

- Seguir depurando el **Proceso de recolección**:
  - Nuevas fuentes Linked Data
  - Procesamiento de lenguaje natural, y otras técnicas de alineamiento
- Añadir nuevos **agentes recolectores:**
  - Para otros tipos de contenidos televisivos: series, eventos deportivos, etc.
  - Información **Geográfica**.
- Modificar la **lógica** de negocio, para que opere directamente sobre el modelo de la BBC y no sobre XMLTV ➔ Conservar URI's.
- Utilizar **frameworks** como Virtuoso Openlink para hacer disponible la información en RDF y SPARQL. Añadir entrada en CKAN, registrar en SIG.MA
- Aplicar principios Linked Data para **usuarios** almacenados en el sistema
- Implementación de **consultas dinámicas** en Linked Data en OntoTV

Futuro

- Los proveedores deben **apostar** por publicar su información utilizando LD:



CAEPIA 2011

FIN

MUCHAS GRACIAS POR LA ATENCIÓN

José Luis Redondo-García, Álvaro E. Prieto, Adolfo Lozano-Tello
Universidad de Extremadura. Escuela Politécnica, Cáceres, Spain
{jluisred, aeprieto, alozano} @unex.es

Software Engineering Group
QUERCUS