Media Meets Semantic Web

- paper from 2009

Definitions

- Semantic Web:
- BBC: British Broadcasting Corporation
- Linked Data:
- DBpedia:
- MusicBrainz:
- NER: Named Entity Recognition

Problem

- British Broadcasting Corporation (BBC) = British public service broadcaster in London
- Large amounts of online content: text, audio, video
- Historically separated into domain specific microsites (food, gardening, sport, etc)
- Not possible to...
    - Find everything, BBC has published to a given subject
    - navigate between BBC domains following a semantic thread (e.g a page about a musician !=> a page with all the programmes that have played that artist)

Objectives

- Better connections and interlinking of existing systems
- Soft transition and reducing impact on existing systems while adding new services to maximize interlinking of domains

1. Develop a service to link all radio and TV programmes with all other data sources in the Linked Data cloud
2. Develop a new music offering
3. Retrofit simple navigational elements (i.e. topic badges)
4. Provide a common set of web scale identifiers to help create equivalency between multiple vocabularies

Interlinking of concepts

- Legacy auto-categorization system: CIS
    - categorize programmes by textual description (brands, locations, people, subjects)
    - difficult to cover every single entity that might be of interest
    - no relations between terms are available (i.e. Beijing and the Beijing Olympics)
    - only internal identifiers, no linking to non-BBC data
    - can be used to interlink between different domains while developing them independently
    - -> if there are mappings between the various vocabularies
- need for a common set of web identifiers: DBpedia!

- DBpedia becomes vocabulary to connect all BBC domains
    - DBpedia Label Lookup: Find most likely matches to a given term, calculate relevance with number of backlinks
    - Context-based Disambiguation: Disambiguate possible matches by clustering them and finding according context in DBpedia
    - i.e. term 'apple' itself is simply a fruit, but in context of 'microsoft', 'google' it becomes 'Apple Inc,'
- (Evaluation)

Interlinking of documents

Identify main actors in a piece of content (*Muddy Boots*):

- parse story body of BBC news URI and use NER system to extract main entities (just text, no semantic meaning or classification)
- an algorithm matches these to possible DBpedia resources
- every possible match of every term is ranked with contextual disambiguation -> this creates mapping of extracted terms to possible counterparts in DBpedia
- identify resources that correspond to 'people' or 'companies' based on present predicates

Content Link Tool

- annotation tool to manually edit metadata
- high quality automated suggestions by Muddy Boots for existing terms or terms from DBpedia