RTÉ Content Discovery*

Christophe Debruyne

Centre for Data Analytics

Insight

Context

The RTÉ Content Discovery project is a collaboration with the Digital Repository of Ireland at the Royal Irish Academy, RTÉ Archives of Raidió Teilifís Éireann (RTÉ) and the Insight Centre at NUI Galway.

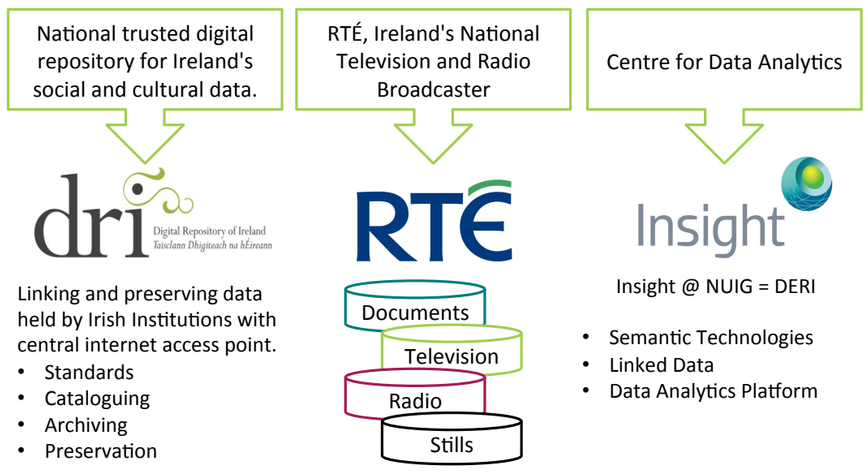

Figure 1: The different partners in the RTÉ Content Discovery project.

Goals and Challenges

The goal of the project is to **discover implicit knowledge** across the different archives *and* the Web of Data to facilitate internal workflows (e.g., search) **for wider reuse and repackaging** of RTÉ's information.

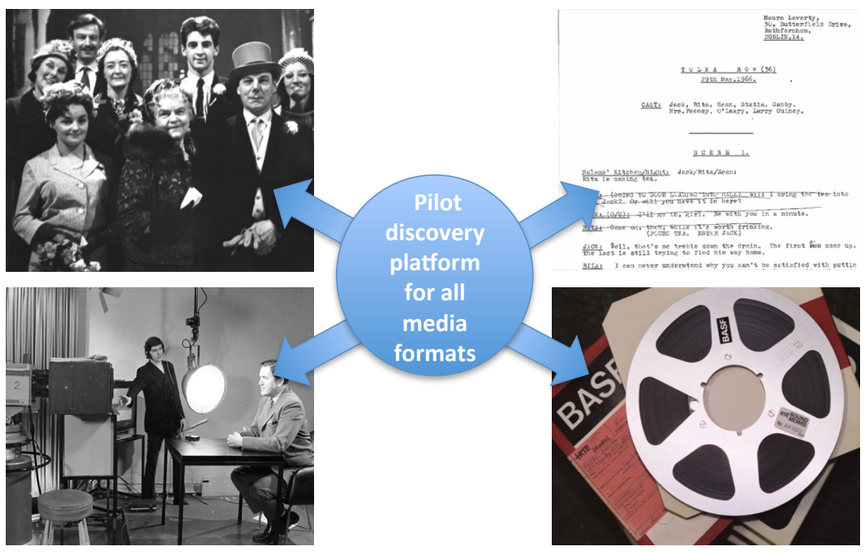

Figure 2: Bringing together content from multiple sources.

To this end, the project aims to create a single access points for the different archives, for which several challenges need to be addressed: heterogeneous databases and different guidelines and practices for each archive; legacy data from previous systems or cataloguing guidelines; English vs. Irish; ...

Tasks

1. **Annotate the data to create RDF using relevant standards, ontologies and vocabularies.**

   Together with a *digital archivist*, who is familiar with cataloguing and best practices in metadata management, samples from the archives related to ''elections'' where chosen, analyzed and cleaned up using OpenRefine.

   RDF from the cleaned data samples were generated by adopting R2RML technologies and creating *mappings* from the data to Semantic Web ontologies such as FOAF, Dublinc Core, EBU Core Owl, etc.

2. **Obtain an integrated view of the different archives by creating links between the RDF representations of RTÉ's archival assets across the different archives.**

   A Linked Data platform was set up in which the RDF representations of the different archives are stored in separate graphs. The creation of links between archives focuses on the entities *keyword*, *place*, and *personality*.

3. **Apply advanced methods for discovering related data for a given subject in external sources such as the Linked Data Cloud.**

   Link discovery is performed by a combination of NLP techniques to discover entities in the assets' descriptions and SILK for declaratively related different RDF datasets and create `owl:sameAs` assertions. The discovered links are stored in a separate graph on the platform.

   Content discovery is focused on supporting the use case of creating an exhibition; the creation of a narrative or story for which content is drawn from the different archives (such as pictures, video and radio fragments) and also investigates adequate tools and visualizations to facilitate content discovery.

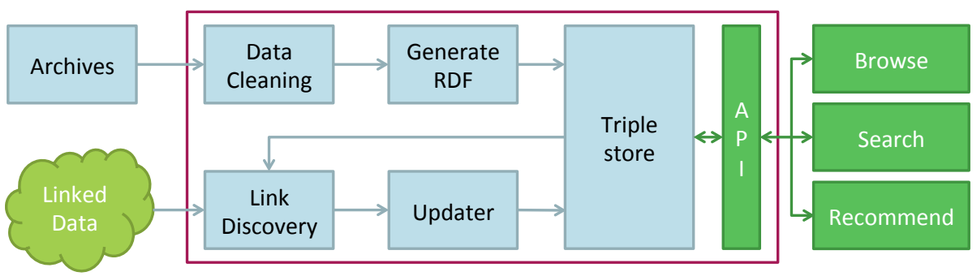

Figure 3: RTÉ Linked Data and Content Discovery Platform.

Methods and Tools

The platform currently holds 6,232 records from the radio archive, 8,335 records from the television archive and 1,673 records from the stills archive. In collaboration with RTÉ, several prototypes for searching and presenting cross-archival searches are being developed and evaluated.

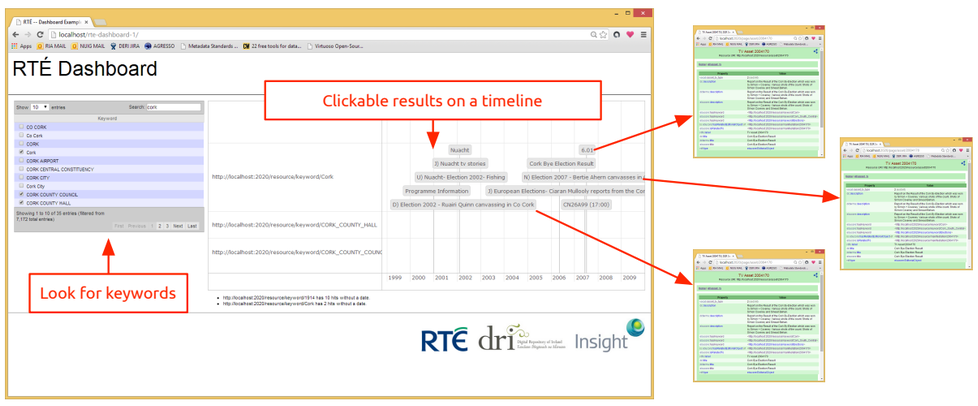

Figure 4: Looking for assets on a timeline.

Taking into account the aforementioned challenges and observations during the cleaning and annotation process, recommendations for each of the archives are formulated to further improve the rich and trusted content in the archives as to facilitate content discovery.

Contact Details

Insight @ NUI Galway `christophe.debruyne@insight-centre.org`
Digital Repository of Ireland `c.debruyne@ria.ie`

*Conducted within the Science Foundation Ireland funded Insight Research Centre (SFI/12/rc/2289).