Digital Repository of Ireland
Taisclann Dhigiteach na hÉireann

Insight

Archiving Events: the Social Repository of Ireland Feasibility Study

Clare Lanigan, Education and Outreach Co-ordinator,
Digital Repository of Ireland

An Roinn Post, Fiontar agus Nuálaíochta
Department of Jobs, Enterprise and Innovation

Ireland's EU Structural Funds
Programmes 2007 - 2013

Co-funded by the Irish Government
and the European Union

HEA HIGHER EDUCATION AUTHORITY
AN tÚDARÁS um ARD-OIDEACHAS

EUROPEAN REGIONAL
DEVELOPMENT FUND

Social Repository of Ireland: Background

Leverage project of the Digital Repository of Ireland (DRI). Project funded by Science Foundation of Ireland (SFI) Technology Innovation Development Award.

Conducted in partnership with Insight Centre for Data Analytics News Lab.

Aim: to test feasibility of building, using & maintaining a social media archiving project for Irish social media.

The Partners

Digital Repository of Ireland:

"The Digital Repository of Ireland is a national trusted digital repository for Ireland's social and cultural data, built by a research consortium of six academic partners working together to deliver the repository, policies, guidelines and training."

(*Source: http://dri.ie/about*)

Insight Centre for Data Analytics News Lab:

"Insight Centre for Data Analytics is a joint initiative of Ireland's leading research centres, working closely with industry partners to develop next-generation data acquisition and analytics solutions for important and diverse application areas."

(*Source: https://www.insight-centre.org/about/mission*)

"Insight Galway's Digital Humanities and Journalism work group is a multidisciplinary effort aiming to bundle research and development activities around news and journalism, media, digital humanities and social sciences with social media and semantic web technologies."

(*Source: http://newslab.insight-centre.org/about/)*

Deliverables

The funding award required the project to complete a set of deliverables in the form of written reports.

These reports involved analysis of archival, business requirements, technical and legal requirements

Deliverable 1: Requirements

A breakdown of project decisions from the beginning.

Established what approaches were adopted and what weren't, and why.

Based on project records – meeting minutes, actions taken, procedures followed.

Covered data models/annotation to be used, metadata standards to be adopted, legal requirements, storage & preservation.

Deliverable 2: Lit Review

Broken into two parts:

1. Comparative analysis of published literature on existing social media archiving projects.
2. Comparative analysis on scope, span and functionality of existing social media collection tools and projects.

Latter divided into data collection tools (e.g. ScraperWiki), modules and libraries (e.g. ARCOMEM) and social media vendors (e.g. Gnip)

Deliverable 3: Vocabularies & Ontologies

Also in two parts:

1. A selective review of existing metadata standards/models used in web archiving & their potential usefulness for projects like Social Repository. Examples include Arizona Model and WAWI.

2. A proposed theoretical ontology and schema design for collecting social media events and reactions. Written by Insight technical team. Describes components of the tool including connections to OA resources such as DbPedia.

Deliverable 4: Conceptual Framework

Detailed description of the structure of the Social Repository data collection tool. Written by Insight tech team.

Includes description of back end components, data sources utilised, front end components, diagrams and workflow charts.

Also includes description of testing done on DRI servers to establish if a Twitter event could be stored and preserved there – established that it could.

Deliverable 5: Legal and Ethical Challenges

Summarised main legal/ethical challenges, including copyright, data protection, libel/defamation, access.

Summarised relevant sections of Twitter Developer Rules with focus on 2011 amendments - their effect on existing projects and subsequent relaxation of the most restrictive aspects.

Findings – Deliverable 1: Requirements

Decision was made to restrict the feasibility study to one social media platform – Twitter.

Insight Labs had the technical infrastructure in place to build and maintain the tool, but not to preserve large amounts of Twitter data.

Solution: DRI could provide access to large scale storage servers used for the repository. The data could also be mapped to DRI's data model for storage.

Findings – Deliverable 1: Requirements

Many discussions about selection and data collection strategy.

It was discussed whether to select events manually or to use an automated collection system e.g. by data bursts, hashtag repetition, etc. A mixture of automatic and manual curation was discussed (the manual aspect being carried out by a digital archivist). In the end it was more cost-effective and useful to set the tool to gather tweet packages automatically relating to activity and hashtag bursts.

Findings – Deliverable 2: Lit Review

Not much up to date published material about social media archiving. Most that exists focuses on web archiving in general, e.g. The British Library UK Web Archive.

Comparative studies of existing projects found that those that were developed by research institutions tended not to be shut down or negatively impacted by Twitter Developer Rules, but also didn't reach as many users as commercial or open source projects.

Findings – Deliverable 3: Vocabularies, Ontologies and Metadata

Established that the metadata fields automatically generated by Twitter for its users' data maps more or less to the metadata standard most commonly used by DRI to date – Qualified Dublin Core (QDC).

As DRI servers would be storing the Social Repository data it would make sense to adopt Dublin Core as a project metadata standard.

DRI also has functionality for other metadata standards – not restricted to QDC.

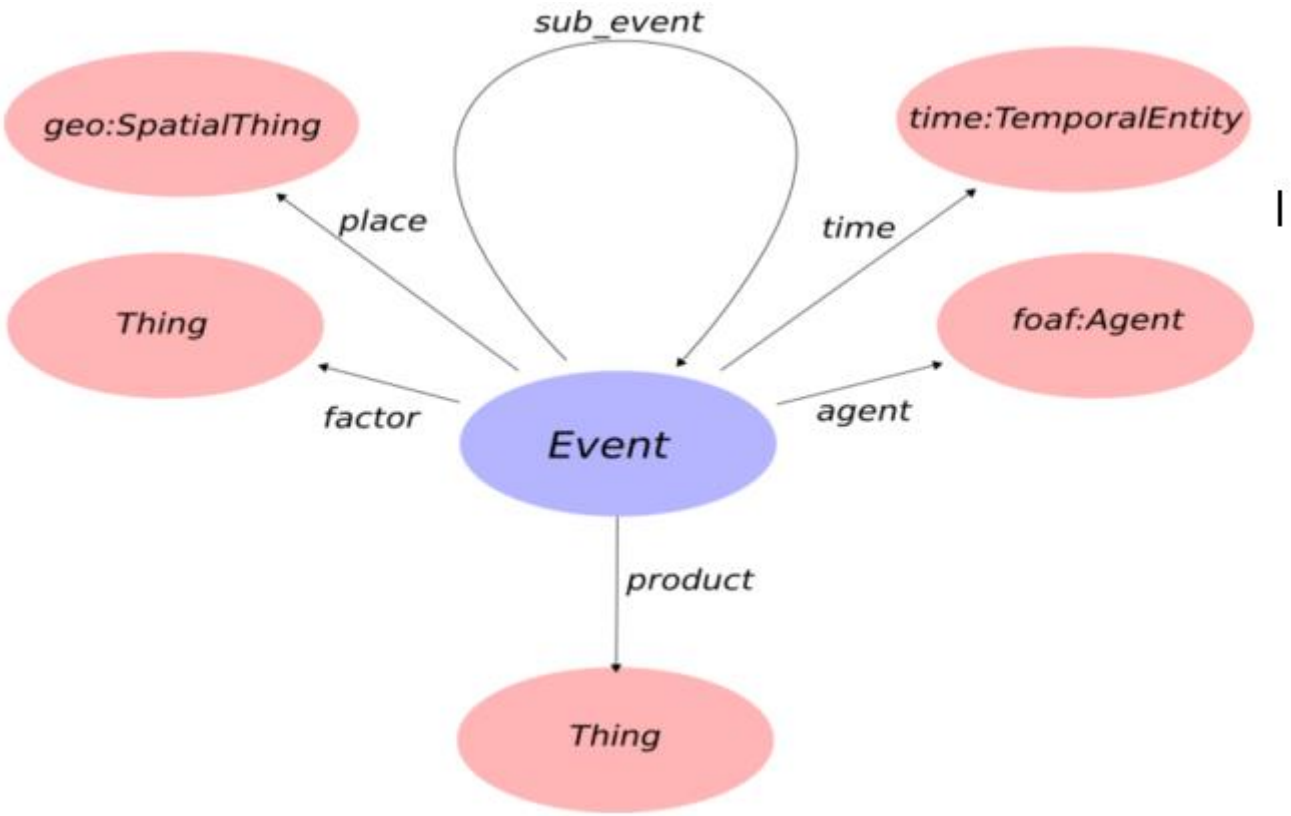
Findings – Vocabularies and Ontologies contd. (Technical)

Report outlines a variety of open source vocabularies that the Social Repository tool could use, including Dbpedia, foaf, Gis-tagging, Geo-tagging and others.

Also includes comprehensive schema visualisations. Next slide provides example:

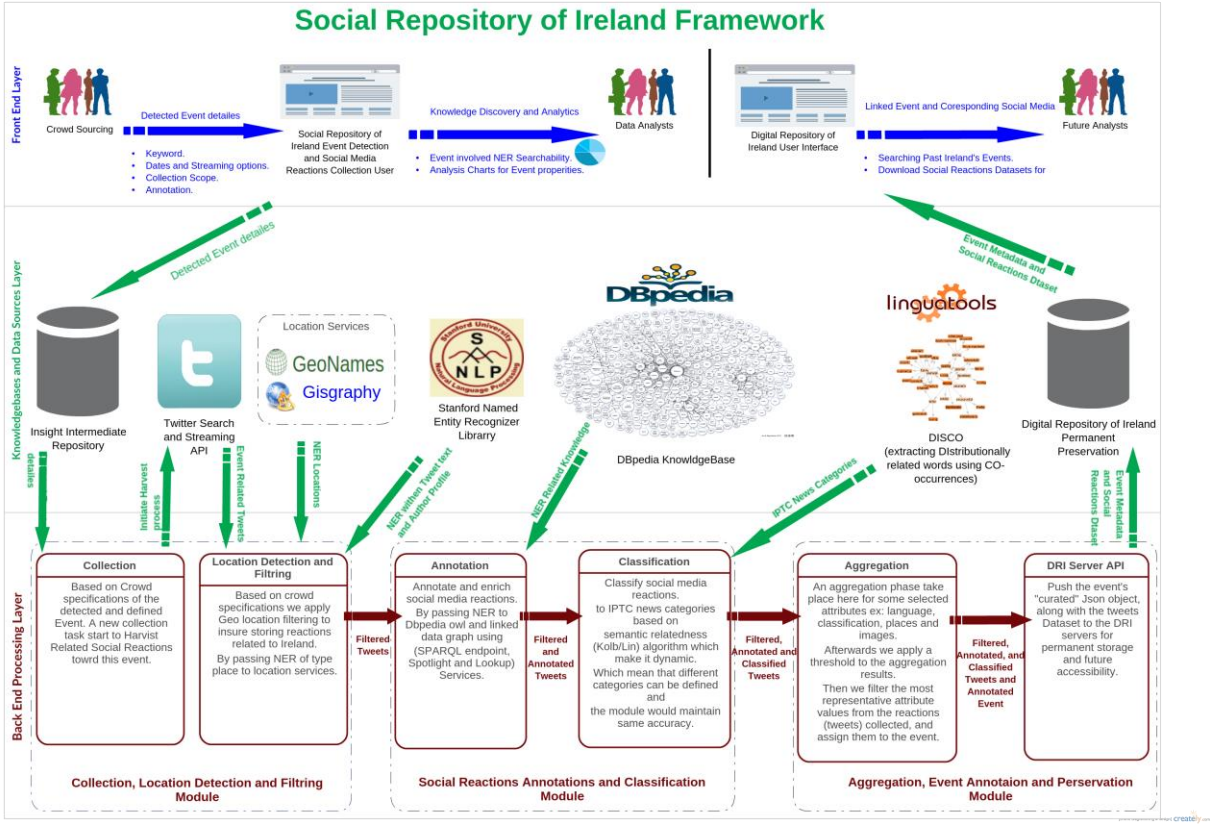
Findings – Vocabularies and Ontologies contd.

Schema – Vocabs used to annotate a Twitter 'event'



Findings – Deliverable 4: Conceptual Framework

Visualisation drawn up by Insight team



Findings – Deliverable 5: Legal and Ethical (Twitter Developer Rules Issues)

Relevant section of Twitter Developer Rules states:

"You may..provide export via non-automated means (e.g., download of spreadsheets or PDF files, or use of a "save as" button) of up to 50,000 public Tweets and/or User Objects per user of your Service, per day."

https://dev.twitter.com/overview/terms/policy#6.Update_Be_a_Good_Partner_to_Twitter

Findings – Deliverable 5: Legal and Ethical (Twitter Developer Rules Issues)

"Twitter's Developer Rules are somewhat vague as to the extent an individual may whitelist tweet data and make it available. The Social Repository has echoed policies adopted by similar tweet collections projects that appear to have maintained their collections without being approached by Twitter for shutdown. These policies involve slight restriction on access and reuse of tweet datasets which fell short of the open-data approach we originally hoped to take. However, this approach does not, in our opinion, place unreasonable restrictions on future users of the Social Repository of Ireland and may be the most workable solution currently available to us."

- Clare Lanigan & Natalie Harrower, DRI, *Social Repository of Ireland Deliverable: Legal and Ethical Report (unpublished)*

Feasibility Study Results

1. Development of **working tool** at Insight News Lab. Not used for 'live' data collection but functionality is fully tested.

2. **Requirements** for archiving data in the form of 'events' (burst of activity around a certain topic or hashtag) established and drawn up.

3. **Technical framework**, vocabularies and ontologies established.

4. Possible **legal restrictions** in Twitter Developer Rules identified and method for **working with** those restrictions established.

The future

SFI TIDA funding allocation only covered feasibility testing. Reports completed but as yet unpublished.

Feasibility has been established. If awards for developing the project become available DRI/Insight would endeavour to apply. As yet unknown if such awards will become available.

Important DRI leverage project – opened up the possibility of social media archiving in the repository at some point in the future.

Further Information

Partners:

Digital Repository of Ireland (http://dri.ie)

Contact: Natalie Harrower @natalieharrower

Clare Lanigan @clarelanigan

Insight Centre for Data Analytics Digital Humanities and Journalism Research Group

(http://newslab.insight-centre.org/)

Contact: Bahareh Heravi @bahareh360