Harmonising Web Feeds for Emergency Management

R. Power[a], <u>C. Wise</u>[a], B. Robinson[a] and G. Squire[a]

[a]*Commonwealth Scientific and Industrial Research Organisation,*
GPO Box 664, Canberra ACT Australia 2601
Email: *catherine.wise@csiro.au*

Abstract:

The Emergency Response Intelligence Capability (ERIC) project is a CSIRO collaboration with the Emergency Management team from the Australian Government Department of Human Services. CSIRO have developed a tool that collects emergency event information from 'live' web feeds published by emergency service agencies and records it in a database. A web feed is a web accessible resource that is updated frequently as new information becomes available from the content provider. This information is made available along with statistical regions and demographics data from the Australian Bureau of Statistics and departmental regional profile data. This allows information for a specific region under investigation to be readily available in the context of the changing status of emergency events. It also provides an archive of the 'live' web feed contents for historical review and analysis, and is a backup when the source is unavailable due to web site outages.

The Emergency Management team monitor emergency events across the country and are responsible for intelligence gathering and situation reporting for the department during emergency events. This information is used to help coordinate the department's response to emergencies, with a focus on delivery of services on behalf of Government for the Australian community. These activities are manual and time consuming, requiring the attention of several staff to obtain and assemble the required information into the appropriate structure and format to produce a situation report. Some of these tasks can be automated allowing emergency coordinators to better utilise their time in the analysis of information rather than finding, collating and formatting it.

In summary, the ERIC tool automatically gathers information from a range of dynamic data sources (the web feeds), stores it in a database, presents the information online using a map, includes a large collection of static data and provides easy access to the information via an interactive map and 'popups'. The tool also provides new functions: the user is notified when new relevant information is available; they can review details of events that have occurred in the past; the user can search for specific events by category, location and source; customised situation reports can be generated for different types of emergency events at specific locations; and a repository of situation reports is maintained. A public version of the tool is available at `http://eric.csiro.au` which has fewer features available, the departmental data is removed and no situation reporting, but it is useful for other agencies to understand the core ERIC features of data integration.

The ERIC tool can be considered a centralised data warehouse replicating the data holdings of the web feed custodians. The contributing sources of emergency event information are heterogeneous in many ways. For example the web feeds publish information using various formats (mostly using structured text), use different conventions for identifying events (using combinations of name, event type, location and severity), publish data updates with different frequency, describe similar events in different and sometimes inconsistent ways, and use the same descriptive labels with different meanings. Maintaining a consistent amalgamated copy of the contributing web feeds is difficult in practice due to these differences in the source systems.

An overview of the various web feeds is presented outlining their structure and content. A detailed description of information recorded from seven web feeds managed by state and territory based fire agencies is analysed noting their differences and similarities. For the period starting early October through to the end of March 2013, there were almost 44,000 individual web feed fire reports recorded describing over 15,500 different fire events, of which 162 were categorised as a *Fire Emergency*. This information is a valuable source of intelligence for situational awareness to emergency managers.

This case study highlights the differing data management practices currently used by the state emergency services and the steps undertaken in ERIC to develop a nationally consistent model of this information.

Keywords: Crisis coordination, disaster management, situation awareness, web feeds, social media

1 INTRODUCTION

The ERIC project is part of the Human Services Delivery Research Alliance (HSDRA); through the Alliance, CSIRO's scientific research capability is used to improve the department's service delivery. The ERIC proto-type tool is being trialled by the Emergency Management team over the 2013/14 disaster season which runs from early October through to the end of March. It is anticipated that the ERIC tool will increase the department's ability to respond to emergencies through the provision of the right information at the right time so that the right decisions can be made in deploying the department's service delivery response. An overview of the ERIC project and tool is described in Power et al. (2013).

2 ERIC OVERVIEW

ERIC's primary interface is the map, allowing the user to navigate the information spatially. The example screen shot of Figure 1 shows the location and expected path of Tropical Cyclone Rusty as reported by the Bureau of Meteorology on 25 February 2013. The warning area has been selected showing the major towns and details of the region under threat. This example highlights some of the important features of ERIC: the recorded information from the various web feeds remains available for historical review and the map provides an intuitive interface for users to readily access information from various agencies.

The ERIC tool consists of three servers, the *Feeder*, the *Web Site* and the *Database*, as illustrated in Figure 2 where each is deployed on different machines for performance reasons. The *Feeder* consists of Python and Java applications which access each feed, seeks new information, standardises the content, and persists the information to the PostgreSQL/PostGIS *Database*. The *Web Site* is a mix of Python and Java deployed on an Apache Web Server. It is responsible for serving pages, saving reports to the database, is a proxy for the tile servers (for background images) and provides a web interface for browser generated client requests. All spatial data is transported to the client web interface as GeoJSON a geographic extension of JavaScript Object Notation (JSON - a text-based format designed for human-readable data interchange).
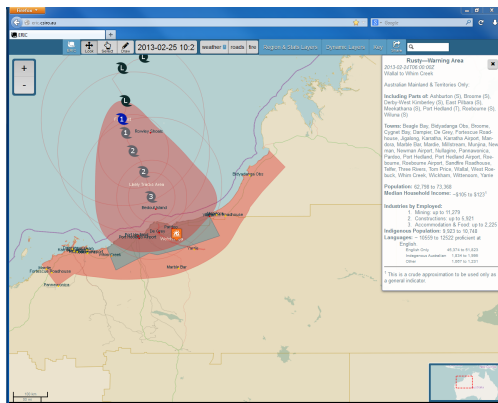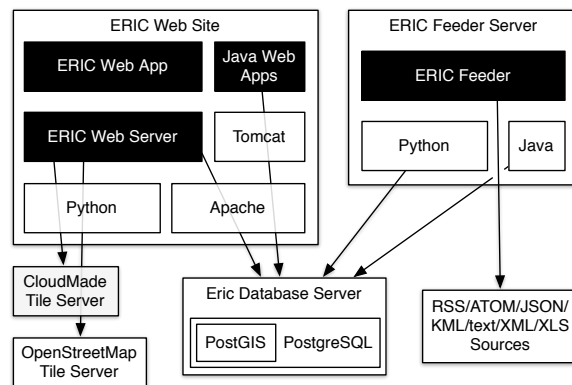

Figure 1. ERIC map interface

Figure 2. System architecture overview

While web feed aggregators exist that provide some of the features of the *Feeder* server[1], they do not support the diverse range of formats currently managed by ERIC. Also, Geospatial Information Infrastructures (GII) and Web 2.0 'mashups', for example GeoCommons (`http://geocommons.com/`), provide some of the ERIC mapping and geospatial features. However, these systems require the data to be available in specific formats and do not support the required situation reporting features.

3 HARMONISING WEB FEEDS

3.1 Web Feed Data Integration

The GeoRSS format, an extension of RSS (Rich Site Summary) to include a location, is commonly used by emergency services agencies to publish information updates about ongoing events that includes details of the geographic location of current incidents, their status and details of the agency's response. These feeds provide content to agency mapping web sites[2], which are intended to provide information about known fires to the

---

[1] See `http://en.wikipedia.org/wiki/Comparison_of_feed_aggregators` for an overview.
[2] For example, the NSW Rural Fire Service (RFS) site: `http://www.rfs.nsw.gov.au/dsp_content.cfm?CAT_ID=683`.

general public. While some agencies, such as the NSW RFS, only publish information about fire related incidents, others such as the ACT Emergency Service Agency (ESA), publish information about a variety of incidents, including fires, hazard reduction, electrical incidents and motor vehicle accidents.

In order to maintain an overview of the status of events occurring around the country, the Human Service's Emergency Management team manually monitor over a dozen such sites to gain a national perspective of incidents underway. This is a time consuming process and places the onus on the team to identify when the emergency situation changes, such as when a new incident occurs or when the status of an existing one changes. The ERIC tool performs these tasks in software: it regularly polls the web feeds for new information, identifying when the situation changes and notifies the user. It also integrates the information into a single web location, allowing the user to maintain a national perspective from a single location. The ERIC tool performs a number of other functions specifically to meet the needs of the Emergency Management team, such as generating situation reports and event searching, see Power et al. (2013).

The purpose of this paper is to focus on the content of the web feeds used by ERIC; a summary is shown in Table 1 indicating the content, data formats and number of feeds used for each. Note that some agencies maintain multiple feeds of varying types.

Table 1. Web Feed Summary

Provider	Content				Format
	Road closures & Traffic incidents	Fires	Weather Warnings	Other incidents	
ACT Emergency Services Agency (ESA)	✓	✓		✓	GeoRSS
Bureau of Meteorology (BOM)			✓		13 GML, 8 RSS, txt, XML
Geoscience Australia (GA)		✓			GeoRSS
QLD 131940	✓				GeoJSON
Qld Dept. of Community Services (DCS)	✓			✓	GeoJSON
NSW Roads and Maritime Services (RMS)	✓				6 GeoJSON
NSW Rural Fire Service (RFS)		✓			GeoRSS
Rural Fire Service Qld (RFSQ)		✓	✓		GeoRSS, XLS
SA Country Fire Service (CFS)		✓			KML
Tas Fire Service (TFS)	✓	✓		✓	GeoRSS, KML
Vic Country Fire Authority (CFA)	✓		✓		2 GeoRSS, RSS
Vic Dept. of Sustainability and Environment (DSE)		✓		✓	2 GeoJSON
Vic Roads	✓				JSON
WA Dept. of Fire and Emergency Services (FES)		✓	✓		GeoRSS

In total, ERIC tracks nine RSS feeds, eight GeoRSS feeds, two Keyhole Markup Language (KML - originally developed for use with Google Earth) sources, 10 GeoJSON sources, and 13 Geography Markup Language (GML) files. The KML and GML files are mostly geometries, containing little descriptive information, while the eight RSS feeds contain no geoencoding and locations are detected using standard phrasings. The RSS and GeoRSS feeds are structured to be human readable whereas the GeoJSON are machine readable and highly structured. An important feature of these web feeds is that they only report the current status of ongoing or recently completed incidents; old incidents are removed and information about them is no longer accessible. Fire Danger Rating Forecast are published by different custodians in four different formats.

Figure 3 shows example content from two web feeds. While XML GeoRSS is used for both, the content is noticeably different: there is little consistency in the tags used; there is no naming convention for events across agencies; the content describing the event sometimes has embedded HTML or semi-structured free text; and different date formats. Across the collection of web feeds in Table 1 there are numerous differences between the information published by the different agencies, even those that describe similar events.

3.2 Overview of Web Feeds

An overview of the web feed content published between September 2012 through to the end of March 2013 is shown in Table 2. Note there is no entry for Qld DCS in Table 2 because information from this feed has only been recorded since April 2013.

Table 2 shows the number of entries in the database for each web feed (the *records* column); the *feed pubs* and *post pubs* information is the number of different publication timestamps recorded for the entire web feed and the posts contained in the web feed content. When these numbers are similar it indicates that the content provider is updating the feed when there is an update to a post. When these numbers are not similar, for example for GA, NSW RFS and Vic CFA, it indicates that when the feed is updated, there are many posts in

```
<item>                                        <entry>
<title>                                       <author>
GRASS AND BUSH FIRE - MITCHELL                <name>Department of Community Safety</name>
</title><link>                                </author><category term="Completed"/>
http://esa.act.gov.au/feeds/                  <content>
currentincidents.xml</link><description>      RFA Type: Traffic Control - Incident Site
Incident: GRASS AND BUSH FIRE - MITCHELL      Date Created: Jul 11 2013  8:55AM
<br />Location: NORWOOD PARK CREMATORIUM,      Property Type:
SANDFORD STREET, MITCHELL<br />               SES Group: Talwood Emergency Services Unit
Status: Out / Completed<br />                 Status: Completed
Suburb: MITCHELL<br />                        </content><id>
Type: GRASS AND BUSH FIRE<br />               urn:uuid:
Agency: ACT Fire<br />                        51bd40ab-4cf4-438a-98e0-3afec3adee72
Incident Number: 030872-10072013<br />        </id><published>
Updated: 10 Jul 2013 18:11:00<br />           2013-07-12T09:42:48+10:00
Time of Call: 10 Jul 2013 17:13:17            </published><title>
</description>                                 6613-20, Traffic Control - Incident Site
<guid>030872-10072013</guid>                  </title>
<pubDate>11 Jul 2013 17:06:00 EST</pubDate>   <updated>2013-07-12T09:42:47</updated>
<georss:point>                                <georss:point>
-35.220086901178 149.133405573183            -28.4862583567276 149.474634535467
</georss:point>                               </georss:point>
</item>                                       </entry>
```

Figure 3. Example GeoRSS Web Feed format and content: ACT ESA (left) and Qld DCS (right).

the content with new information because they have different publication timestamps. The *events* column in Table 2 is the number of different events described by the web feed while *description* is the average length of the description text. The next four columns record the number of different URL links, categories, points and areas. The last column is the average time from the first to the last post for the events, recorded as hours:minutes or by days (*D*).

Table 2. Web Feed Summary. Fire web feeds indicated with a [†].

Provider	records	feed pubs	post pubs	events	description	link	category	point	area	duration
ACT ESA[†]	6760	5147	5150	1354	344	145	8	1443	0	2:29
BOM	3939	3246	3655	3640	2066	128	10	500	445	0:03
GA	118696	1291	13195	87861	144	0	1	89797	0	0:01
QLD 131940	17711	7425	7425	10158	239	0	6	6778	0	2.4 D
NSW RMS	55600	20627	20627	14882	830	0	13	8773	0	4.6 D
NSW RFS[†]	21541	7767	15707	11317	281	2	9	7386	1256	3:20
RFS Qld[†]	18315	18315	14223	13724	253	13977	6	12194	0	2:04
SA CFS[†]	11490	7513	7513	9632	99	0	6	5129	0	1.8 D
Tas FS[†]	22542	11956	8298	5063	379	5070	6	4003	0	7:14
Vic CFA[†]	148110	23225	78178	45804	366	1442	5	39675	0	2:02
Vic DSE	143188	22390	22390	43860	261	0	26	37837	0	2:01
Vic Roads	50	4	4	3	4	0	1	50	45	0:00
WA FES[†]	1611	1611	1295	1281	2862	1295	8	1323	0	9:16
Total	569553	130517	197660	248579	8128	22059	105	214888	1746	17:53
Combined	569553	130399	192512	248579	342	22059	69	178730	1746	11:52

The *Total* row is the sum of the above column whereas the *Combined* row is the sum accumulated without regard to the feed. For example, the *Total* for the *category* is 105 however *Combined* is 69: there are 69 different categories from all feeds, indicating that the same category value is used by different web feeds.

3.3 Harmonised Model

Table 1 and Figure 3 show that various agencies around Australia provide detailed information of assorted events using different data formats and structures. The web feed content is conceptually the same for each: a feed has a URL endpoint which publishes updates as *content*. The *content* consists of one or more entries, also referred to as *posts*, which describe the individual events. The *post* consists of: a unique identifier allowing information about the same event to be tracked; another timestamp since the web feed content may contain a mix of old and new posts; the post content; and the location of the event as a geographic coordinate. This common model is shown in Figure 4 which also shows some extra information extracted from the *description*: the event *category*, a *link* to further information from the content provider and the *area* impacted by the event.
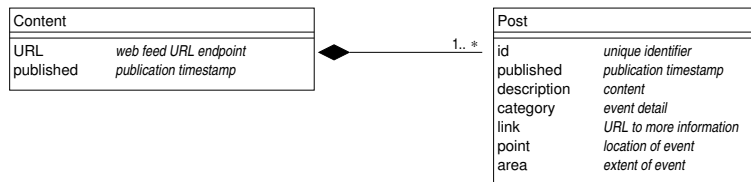
Content	
URL	*web feed URL endpoint*
published	*publication timestamp*

Post	
id	*unique identifier*
published	*publication timestamp*
description	*content*
category	*event detail*
link	*URL to more information*
point	*location of event*
area	*extent of event*

Figure 4. Common Web Feed Model

3.4 Summary

The content from the seven fire web feeds, indicated by a † in Table 2, for the period 1 October 2012 – 31 March 2013 were examined to determine the total number of different events reported and their alert levels. This information is shown for all the target web feeds combined, noting the total incidents per day by alert level in Figure 5(a) and the combined incidents by each state based agency in Figure 5(b).
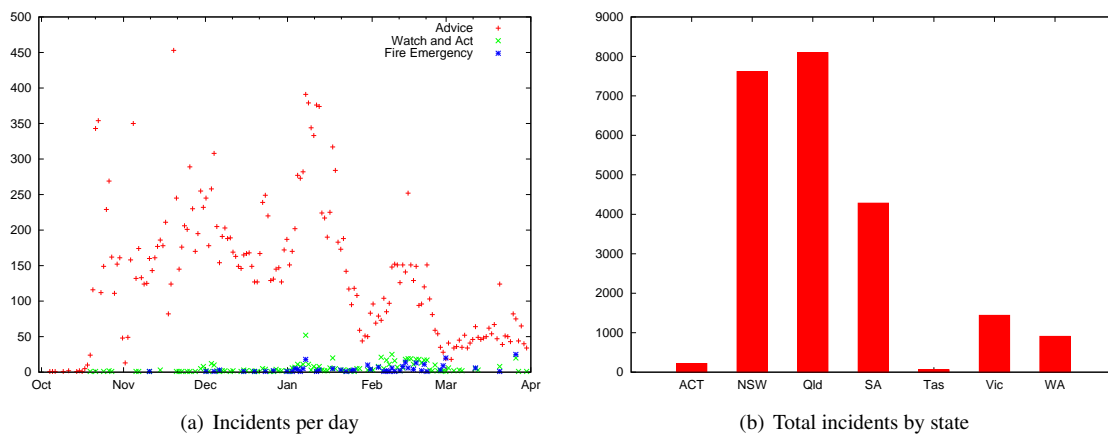

(a) Incidents per day	(b) Total incidents by state

Figure 5. Alert Levels

Figure 5(a) illustrates the rise and fall of the season, with numerous sporadic fires occurring early in the season and fewer, more intense fires later on. This is especially true in February and tapers off quickly in March. Also, the higher density of *Watch and Act* and *Fire Emergency* reports aligns with the peaks in the number of *Advice* reports issued. The difference in how the agencies report is indicated by Figure 5(b), where it can be seen that NSW and Queensland are far ahead.

4 HARMONISING IN PRACTICE

As noted above, the *Feeder* is responsible for accessing each feed, identifying when new information is available, standardising the formatting and content and persisting the information to the PostgreSQL/PostGIS database. The *Feeder* server regularly checks the web feeds for new *content*, currently every 10 minutes, and only new updated *content* is recorded in the *Database*. Since the web feed *content* consists of a collection of many *posts*, an update may consist of a collection of *posts* where some have been changed and some not. This is managed by the *Feeder* server generating an SHA-1 hash value of the *post*: when the hash value changes, the *post* has been updated and it is recorded in the database. This process also allows the ERIC tool to examine the *post* content further to identify when the status of events has changed, for example when a fire event escalates to an *Emergency Warning*.

For the web feeds noted in Table 2 above, there have been a number of issues with this process over the first year of operation. These are described below.

4.1 Model Based Approach

The *Feeder* employs a model based approach. The model of Figure 4 is the structure resulting from *Feeder* processing. For example, when looking for the category of a KML source, the *Feeder* identifies the 'styleUrl' tag, while in JSON the appropriate dictionary entry is one of the keys 'Type', 'incidentType' or 'category'.

As shown in Figure 3, the web feed content that describes events varies. In general, the information required can be summarised by answering the five '*Ws*': who (the web feed provider), what (the event type or category), when (how recently was the information published), where (the point location or area affected by the fire) and why (what is new about this post for it to have been published). The answers to these questions can be readily obtained from the web feed content, although some extra processing of the post *description* is required.

As noted above, the various feeds describe different kinds of events. For fire events, there is extra information characterised as the fire 'alert level' encoded as part of the event *category*. The various fire agencies, indicated by a † in Table 2, report this information in different ways. For the period 1 October 2012 – 31 March 2013, there were 21 different categories of events derived from the content of these feeds, ranging from hazard reduction, permitted burns, car incidents, false alarms, electrical fires and so on. Of these, the information of interest is the fire 'alert level', the fire warnings to the community. In general, the progression of alert levels[3] for fires are *Advice*: an indication that a fire has started and there is no immediate danger; *Watch and Act*: a heightened level of threat where preparatory action by the community is required; and *Fire Emergency*: the highest alert level where there is immediate threat to the community.

The different fire agencies noted in Table 2 report this information in different ways. Of these seven web feeds, four (NSW RFS, RFS Qld, Tas FS, Vic CFA) use fire warnings in the three alert level categories noted above and WA FES mentions the alert level in the event title. The remaining two (ACT ESA and SA CFS) do not directly state the alert level, but instead provide detailed descriptions of the type of fire, such as 'grass fire'. In these two cases, the alert levels are mapped as an *Advice* for all reported fires.

While the simple model of Figure 3 goes some way to provide a common conceptual model of the different web feeds, further investigation of the various web feeds is required to improve this model. The mappings to the model are currently not declarative, they're captured as feeder-specific python code, which makes them difficult to extend, re-use or compose. It also makes it difficult to describe the feeds in a flexible or generic fashion, to allow for more intelligent processing that could allow for some changes and/or errors in the feeds.

4.2 Internal Web Feed Changes

There have been occasions where there have been small changes in the web feed structure requiring adjustment to the *Feeder*. The *Feeder* evolved into robust software through iterations so that such changes would not cause errors in processing; initially, there were occasions where such changes went unnoticed for weeks and we subsequently lost information. Parsing is done using third party libraries for standard formats (KML, RSS, JSON etc) and the *Feeder* handles most errors gracefully. It is however brittle in the face of changes to feeds.

The most common issue encountered has been the inclusion of special characters (such as decorative bullet points), which are simply removed. On occasion, the spatial location is missing and Nominatim, a geo-coding service which is part of the OpenStreetMap project, is used to to infer a location from text within the *post*. Also, the structure of the feed *content* has been known to change. For example, the JSON dictionary keys for certain fields have been re-named, for example changing 'main-category' to 'mainCategory'.

While the changes to the *Feeder* have to date been minimal and only require small effort to rectify, the combined loss of information is in the order of several weeks. As a result, a new *Monitor* software component has been introduced which periodically checks, for each feed, the timestamp of the most recent record written. Based on previous data processing experience, each web feed is expected to have new *post* entries written to the database at least every 30 minutes. When no entries have been seen for this period of time, an email is sent to the development team.

4.3 Web Feed Access

The monitor has also proved useful to indicate when there are other issue with the software infrastructure, such as problems with virtual machines and Internet connectivity. Two categories of web feeds are not always active: most states turn off their fire danger rating feeds over winter and the BOM only provide information about cyclones when there is an active cyclone. The BOM provide this information as a file in a folder on an FTP server. If there are no active cyclones, the folder is empty. While we can tell if the entire server is down, we cannot tell if there is a problem with the just the cyclone data. The other feeds are server based and so connection issues are indicated using HTTP error codes, for example a response code of 404 or similar, but we can not tell if the data is missing, moved, or if the whole server is down.

---

[3]These descriptions were taken from: `http://www.rfs.nsw.gov.au/dsp_content.cfm?cat_id=2729`.

4.4 Integrating New Web Feeds

A new web feed can be introduced with minimal effort by the current development team, however this requires detailed knowledge of the feeder code base. The main issue in terms of maintenance is the software engineering expertise required to obtain the knowledge. Also, the effort required will depend on how different the new feed is to existing feeds. In summary, minor changes need to be made to the database schema to include a reference for the new feed and *Feeder* code written to accommodate any unusual elements of the feed content.

For example, it has been straight forward to integrate a new GeoRSS feed. The main effort has been in harmonising the categories. Some feeds don't have links, or at least not permanent links, use agencies specific and non standards categories to describe, for example, fire events, such as '/other/closed', and some entries don't have spatial information.

4.5 Access Restrictions

The feeds mostly have open licenses, see `http://creativecommons.org/licences` for an overview, allowing the ERIC tool to show the information they publish. This is not true, however, for the Victorian Country Fire Association, and the Victorian Government Department of Environment and Primary Industries, which do not allow any republishing. For the purpose of this research they have been included in a version of ERIC available to select Department personnel, however they are not available on the public site, leaving a noticeable gap in Victoria for current fires and fire danger forecasts. Harmonizing data where licenses may be unclear, or restrictive, is likely to be an ongoing issue, especially dealing with diverse agencies.

5 CONCLUSION

In this paper we have presented ERIC, a tool that collects emergency event information published via web feeds, harmonises the different data formats used, stores the information in a database for later use, identifies when new information is available or the status of known events changes and includes other datasets from various government agencies. The map based interface allows a user to readily access this information and the tool is currently being trialled by the Emergency Management team. Our investigation has focused on fire information published by seven agencies from around Australia over the 2012/13 disaster season. We found that there are very different styles and approaches to fire reporting between the different agencies.

Further work is planned to include new data sources, refine the common data model to expand the information describing events and promote ERIC to other government agencies. We are also exploring the use of social media where our initial investigations indicate that information is published on Twitter before updating the web feed, is reported more frequently and contains more specific event information, see Power et al. (2013).

The web feed *Feeder* is currently bespoke software where each feed requires its own code for processing. This architecture has been suitable for developing a prototype. A better solution is a model driven approach where a conceptual model of the information to be integrated is defined and the individual feeds are mapped to the common model. The investigation undertaken to prepare this paper was partly motivated to further understand the structure of the individual web feeds so that an enhanced common model could be developed.

REFERENCES

Power, R., J. Dickinson, C. Wise, and B. Robinson (2013, May). Emergency Response Intelligence Capability: Improving Situational Awareness in the Australian Government Department of Human Services. In *Australian and New Zealand Disaster and Emergency Management Conference Proceedings*, PO Box 29, Nerang QLD 4211, Australia, pp. 191–209. AST Management Pty Ltd. `http://anzdmc.com.au/archives/disaster.php`.

Power, R., B. Robinson, and C. Wise (2013). Comparing web feeds and tweets for emergency management. In *Proceedings of the 22nd international conference on World Wide Web companion*, WWW '13 Companion, Republic and Canton of Geneva, Switzerland, pp. 1007–1010. International World Wide Web Conferences Steering Committee.