DELIVERABLE SUBMISSION SHEET

To: Susan Fraser *(Project Officer)*

EUROPEAN COMMISSION
Directorate-General Information Society and Media
EUFO 1165A
L-2920 Luxembourg

From:

Project acronym: PHEME Project number: 611233

Project manager: Kalina Bontcheva

Project coordinator The University of Sheffield (USFD)

The following deliverable:

Deliverable title: Evaluation results: validation and analysis

Deliverable number: D8.4

Deliverable date: 31 March 2017

Partners responsible: Schweizerische Radio-und Fernsehgesellschaft Association (SWI)

Status: ☒ Public ☐ Restricted ☐ Confidential

is now complete. ☒ It is available for your inspection.
☒ Relevant descriptive documents are attached.

The deliverable is:

☒ a document
☐ a Website (URL: .....)
☐ software (.....)
☐ an event
☐ other (.....)

Sent to Project Officer:	Sent to functional mail box:	On date:
Susan.Fraser@ec.europa.eu	*CNECT-ICT*-611233*@ec.europa.eu*	04 April 2017

FP7-ICT Strategic Targeted Research Project PHEME (No. 611233)

Computing Veracity Across Media, Languages, and Social Networks

D8.4 Evaluation Results: Validation and Analysis

Peter Tolmie & Rob Procter (University of Warwick)

Christian Burger & Geraldine Wong Sak Hoi (SWI swissinfo.ch)

David Losada (iHUB)

Abstract

FP7-ICT Strategic Targeted Research Project PHEME (No. 611233)

Deliverable 8.4 (WP 8)

This deliverable describes the activities undertaken to evaluate the PHEME Journalist Dashboard that was developed over the course of Work Package 8. It presents findings from each of three formative evaluations that were conducted over the course of the final year of the project. In these evaluations ways in which the dashboard would clearly be able to support journalistic work were identified. At the same time each evaluation exercise also revealed certain issues that were then tackled in development such that, for each new set of tests, the dashboard was improved from its prior version. The deliverable also presents a comprehensive set of findings from the final summative evaluation of the dashboard and a parallel evaluation of the Hercule fact-checking dashboard. These findings amount to a statement of what the technical work in PHEME directed towards the requirements identified in Work Package 8 was finally able to accomplish. We conclude with some reflections upon what would be needed to take the dashboard further into an actual working tool embedded in journalistic workflows.

Keyword list: evaluation, journalistic work requirements, PHEME dashboard

Nature: Report

Dissemination: PU

Contractual date of delivery: 31 March 2017
Actual date of delivery: 04 April 2017

Research partially supported by the European Union under the Information and Communication Technologies (ICT) theme of the 7th Framework Programme, grant PHEME (611233)

PHEME Consortium

This document is part of the PHEME research project (No. 611233), partially funded by the FP7-ICT Programme.

University of Sheffield
Department of Computer Science
Regent Court, 211 Portobello St.
Sheffield S1 4DP, UK
Tel: +44 114 222 1930
Fax: +44 114 222 1810
Contact person: Kalina Bontcheva
E-mail: K.Bontcheva@dcs.shef.ac.uk

Universitaet des Saarlandes
Language Technology Lab
Campus
D-66041 Saarbrücken
Germany
Contact person: Thierry Declerck
E-mail: declerck@dfki.de

MODUL University Vienna GMBH
Am Kahlenberg 1
1190 Wien
Austria
Contact person: Arno Scharl
E-mail: scharl@modul.ac.at

Ontotext AD
Polygraphia Office Center fl.4,
47A Tsarigradsko Shosse,
Sofia 1504, Bulgaria
Contact person: Georgi Georgiev
E-mail: georgiev@ontotext.com

ATOS Spain SA
Calle de Albarracin 25
28037 Madrid
Spain
Contact person: Tomás Pariente Lobo
E-mail: tomas.parientelobo@atos.net

King's College London
Strand
WC2R 2LS London
United Kingdom
Contact person: Robert Stewart
E-mail: robert.stewart@kcl.ac.uk

iHub Ltd.
NGONG, Road Bishop Magua Building
4th floor
00200 Nairobi
Kenya
Contact person: Rob Baker
E-mail: robbaker@ushahidi.com

SwissInfo.ch
Giacomettistrasse 3
3000 Bern
Switzerland
Contact person: Peter Schibli
E-mail: Peter.Schibli@swissinfo.ch

The University of Warwick
Kirby Corner Road
University House
CV4 8UW Coventry
United Kingdom
Contact person: Rob Procter
E-mail: Rob.Procter@warwick.ac.uk

Executive Summary

In this deliverable we present the results for the evaluation of the PHEME journalist dashboard.

We begin by describing the overall evaluation process of the PHEME dashboard, including a summary of the three formative evaluations (the first two have been reported in deliverable D8.3), which are then followed by a detailed account of the comprehensive methodology adopted for the final, summative evaluation.

We then present the results of evaluations, with a particular focus on the summative evaluation, which was designed to assess each of the principal features, the experience of using the dashboard in its own right, and how well the dashboard compared to other resources journalists are used to using. The summative evaluation included a parallel evaluation of the Hercule fact-checking dashboard developed by the Ontotext partner, thereby enabling a comparison of the two dashboards.

We conclude by discussing the main implications of these results for the future development of the two dashboards. The results provide clear evidence that there are a number of areas where further work would be required to bring both dashboards up to a production quality state, where they would be ready for newsroom deployment; most notable is the need to provide for flexibility in use so as to be able to accommodate variations in newsroom verification practices. Nevertheless, the results also confirm that the majority of the goals in Work Package 8 have been accomplished.

Contents

1 Introduction

Work Package 8 in the PHEME project has been dedicated to the initial development of the use case for journalism, followed by the active creation of a journalism dashboard. There is a natural fit between the interests of PHEME in the detection of rumours, together with the assessment of veracity in social media content, and the work of journalists where the requirement for rapid verification of User-Generated Content is already well-documented (Calcutt & Hammond, 2011; Harkin et al, 2012; Singer, 2015). Initial effort was therefore focused upon understanding the exact nature of the requirements present in journalistic work and this was explored through several in-depth ethnographic studies (see PHEME deliverable D8.1, 2015).

The ethnographic work was then drawn upon to begin the process of creating a working prototype dashboard that could be iteratively tested and evaluated with journalists and then refined. The journalism use cases and outcomes of some of the initial evaluations are already reported in Deliverable D8.3 (2016). In this report we finalise the description of the Journalism Dashboard evaluation exercises undertaken in PHEME and examine their outcomes. In the final, summative evaluation journalists also looked at the fact-checking dashboard, Hercule, which was developed in the latter stages of the project by Ontotext. Findings from that evaluation are also reported here.

As will be seen, whilst there are a number of areas where further technical work would be required to bring both dashboards up to a fully integrated working state for the generation of real-world news content, the majority of the goals in Work Package 8 were accomplished. It should also be noted that the creation of a fully functioning Journalism Dashboard pulled upon technical expertise from across the consortium and demanded intensive cross-partner collaboration and coordination, which may comfortably be counted as another of the project's successes. The technical backdrop to the evaluations reported here is separately reported in D8.3.1 and for a fuller description of the dashboards being evaluated this report should be read alongside of that deliverable.

The report is structured around an initial outline of the work package, followed by a description of the various evaluation exercises undertaken. A results and analysis section then explores the outcomes of the various evaluations, their implications, how these were acted upon, and what can be said about the final state achieved. We conclude with some reflections upon what would be required to develop the technology further. At the end of the report some of the documentation related to the various evaluation exercises is also included as an appendix.

2 Intended Audience

This document details the work undertaken in order to ensure that Task 8.4 of Work Package 8 of the PHEME project met its necessary objectives within that work package. In order to ensure a quality outcome internal procedures were put in place for this deliverable to be peer reviewed and verified prior to submission.

As such, this document is intended for all actors involved in the PHEME project.

3 Work Package 8 Overview

Task 8.1 Requirements gathering, use case design and interface mock-ups

Duration: January 2014 – December 2014 (prolonged due to adapted methodology)

Task 8.2 Journalism Corpus Collection and Annotation

Duration: January 2014 – June 2015

Task 8.3 Open-source digital journalism showcase

Duration January 2014 – March 2017

Task 8.4 Iterative Evaluation

Duration August 2014 – March 2017

4 The Evaluations

The focus of Task 8.4 in the work-package has been upon developing and applying an appropriate set of evaluation activities for the PHEME Journalist Dashboard. These were organised iteratively throughout the whole of 2016 and up until March 2017. There were 4 key checkpoints within the process, amounting to 3 formative evaluations in January, April and September 2016, and a fourth summative evaluation that was initially planned for December 2016 but postponed to February/March 2017 to allow for completion of the prototype. Between these relative structured checkpoints there were ongoing evaluation activities addressed to the functionality of specific features, as and when the need arose.

The concepts of formative and summative evaluation originated in projects designed to bring about changes in educational practice (Scriven, 1967). In technology and computing a formative evaluation is one that is undertaken during the course of a technology's development in order to inform the further design and refinement of specific elements. A summative evaluation is one that is undertaken once the technology is (notionally) finished (or at least at the end of that particular development cycle). Here the goal is to evaluate how the technology performs overall, usually with an eye to whether the technology should be taken up and used, productised, sent back for re-design, dropped entirely, etc. (Koenemann-Belliveau et al, 1994). It should also be noted that both approaches in the case of this project adhere to the principles of what is called ‘situated evaluation’ (Twidale et al, 1994). The point of situated evaluation is effectively threefold. That is: 1) it involves evaluating use in situ, and with the use of in situ observation (as opposed to evaluation through experiment in an experimental setting); 2) use of the system is evaluated by having it situated within real-world workflows and the conduct of real-world activities; 3) the effectiveness of the system is evaluated with regard to ‘its ability to be situated or embedded in the work of a setting’ (Crabtree et al, 2012).

All of the evaluation exercises involved a core of professional newsroom journalists from within

SWI. The final summative evaluation also involved a journalist from one of SWI's partner organisations. All of these journalists were drawing upon social media for their work on a daily basis and most were either currently active on newsdesks or had been in the recent past. Full demographic detail can be found in Section 5.2.

On top of this, interim evaluation was undertaken by members of the project team within SWI so that individual features could be quickly tested as soon as they were ready and then refined without needing to wait for a more formal evaluation exercise. In all instances the reference point for evaluation was real-world journalistic work practices and associated workflows and the methodology was directed towards testing the effectiveness of the technology for supporting journalists in real-world work situations.

The final evaluation was additionally structured in such a way as to enable us to assess how PHEME technology compared to other tools currently being used by journalists and to what extent journalists would see it as an important addition to their armoury. It was also designed to assess how automated features such as veracity assessment, factual accuracy, and controversiality measures would compare to existing largely manual forms of assessment in journalistic work. A further focus of the evaluations was the performance of the dashboard in relation to newsroom concerns about timeliness in the delivery of results. Against all of these concerns the final evaluation also sought to get a measure of the quality of experience of using the dashboard for the journalists involved.

The formative evaluations followed a situated evaluation model and concentrated on observing journalists attempting to use the dashboard in the context of a typical workflow. The summative evaluation also made use of a range of questionnaires to collect active user ratings of various features and concepts. Instruction sheets used in the evaluations and the final questionnaires are included with this report as an appendix.

Throughout all of the evaluations each journalist was asked to use the tool in a separate session. This meant that members of the project team could observe each evaluation in situ, cultivate feedback, and pose additional questions where appropriate. For the larger part members of the teams at Warwick and at SwissInfo managed the sessions. Other project partners joined the sessions via Skype or Google hangouts and were given access to the desktops being used by the participants. For technical partners in particular this provided the opportunity for clarification of certain features as the need arose, not to mention a to-hand channel for alerting people to arising issues and discussing potential solutions.

Video and audio recordings were made of each of the evaluation sessions as well as screen captures so that interaction both with and around the dashboard could be captured and analysed. In all cases consent was sought and received from the journalists prior to any recordings being made and work-up of the materials has been systematically anonymised. Analysis of the evaluation outcomes was commensurate with the ethnographic approaches outlined in D8.1 and focused upon how the technology was reasoned about by the journalists as they encountered it. As will be seen in the evaluation results the analysis is centred upon three key concerns:

- how usable the technology is in practice;
- how intelligible the concepts underlying the technology and its organisation are found to be;
- and how effectively the technology might be embedded within actual working practice.

4.1 Formative Evaluation 1

As was reported previously in D8.3, the first formative evaluation took place in January 2016 and involved 2 journalists from SwissInfowho were experienced users of social media in a newsroom context. In separate sessions each of the journalists used a prototype of the PHEME journalist dashboard to access canned data from a pre-annotated dataset related to the Germanwings crash in March 2015. They were instructed to use the dashboard as though it was part of their regular workflow and they were going to construct a story, drawing information from the dashboard alongside of other routinely used tools, such as the newswires, Google, other news sites, and Microsoft word for compiling an initial draft. Each session lasted about two hours and video recordings were made of their screens and their interactions with their computers. They were also asked to vocalize their reasoning and impressions throughout the session and audio recordings were made as well. The **primary goal** in this first evaluation was to get a sense of how the tool would be interleaved with their existing workflow and what kinds of features they would focus upon as potential resources.

4.2 Formative Evaluation 2

The second formative evaluation was also reported in D8.3. This evaluation was conducted in April 2016, using the same journalists as we had used in the first evaluation exercise so that they could make direct comparison. The **primary goal** of this evaluation was to conduct a similar exercise to the first evaluation, but this time drawing upon live data. On this occasion technical difficulties and unforeseen operational constraints prevented the journalists from pursuing exactly the same steps as they had previously. Nonetheless, the evaluation surfaced a number of important issues to be resolved and was a critical staging point in the definition of requirements and their relative priorities. Outcomes of this evaluation in particular shaped much of the technical development throughout the rest of the project.

4.3 Formative Evaluation 3

A third formative evaluation was conducted in September 2016. This evaluation used the same two journalists as before, but this time a third, German-speaking journalist was also invited to undertake the same set of tasks. Once again the instructions focused upon getting the journalists to pursue a standard story construction workflow, using the dashboard alongside of other resources to track down potential story leads, assess the veracity of various facts, and identify potential content. As with the second evaluation the intention was to draw upon live data. In view of the anticipated later release of a German version of the dashboard, the German journalist was also able to make use of an interface where most of the terms had been translated into German, although the tweets streamed into the dashboard continued to be in English. The **primary objective** within this evaluation was to test the range of revisions that had been undertaken after the prior evaluation so as to identify what further work would be required to take the technical development through to the project end and the final, summative evaluation. With this in mind, the evaluation on this occasion also provided opportunity for the journalists to test out specific features and their functionality so as to provide targeted feedback on the organisation of the interface and the intelligibility and usability of the various components.

4.4 Summative Evaluation

A final, summative evaluation of both the PHEME journalist dashboard and the associated fact-checking dashboard developed by Ontotext, Hercule, was conducted across February and March 2017. The same two journalists who had been involved throughout the evaluation process at SwissInfo again participated in this exercise. The German-speaking journalist used in the third formative evaluation also took part. A further journalist from SwissInfo was invited to participate as well so as to provide the possibility of having it evaluated by someone who had not previously seen the tool. On top of this a journalist from one of SwissInfo's partner organisations took part so as to give a response from someone who was using social media in a different workflow in a different organisation. This made a total of five journalists providing feedback across a range of different tasks.

As outlined above, the purpose of a summative evaluation is somewhat different to a formative evaluation. At this point the assumption is that no significant further refinement or change of the technology will take place within the current phase of development. A summative evaluation therefore sets out to test the readiness of a technology for use in its current state and to assess what enhancements might be useful in future rounds of development. With this in mind the tasks within the final evaluation were more varied and far-reaching. As on previous occasions one of the sets of tasks sought to assess the usability of both dashboards within the context of each journalist's regular workflow. For both dashboards this was undertaken using live data as selected by the journalists themselves. On top of this, however, the journalists were also asked to work systematically through the principal features of each dashboard and to provide feedback both verbally and via questionnaires regarding three principal concerns:

- the intelligibility of the features and associated concepts;
- the actual usability of the features as a matter of practice;
- and the extent to which the features might be useful additions to the current suite of tools on offer.

For the PHEME journalist dashboard this exercise was conducted using previously collected tweets relating to the Charlie Hebdo attacks in France in 2015 rather than live data. This was to ensure that each journalist was working with a comparable dataset with some tasks being set in relation to known and stable outcomes. For Hercule this part of the evaluation also used live data as the range of tasks had a more restricted scope and the possible outcomes were more predictable. A third block of tasks, administered after the principal situated evaluation exercise had taken place, asked the journalists to provide some background demographic data and their overall impressions of each dashboard.

All the situated evaluation exercises were recorded in a variety of ways, including screen capture, audio recording, and video recording, with each journalist attending a different session and being invited to provide verbal commentary throughout.

As results from all but the third formative evaluation have already been reported in a prior deliverable, this report will concentrate primarily upon the results of the summative evaluation, across each of the blocks of tasks (feature-assessment; workflow; and experience).

5 Results & Analysis

In this section of the report we will be presenting the results of the various evaluation exercises undertaken in PHEME between January 2016 and March 2017. In the first part of the section we will summarise the outcomes of the formative evaluations, many of which have already been reported in greater detail in D8.3. The second part of the section will concentrate exclusively upon the outcomes of the final, summative evaluation. The basic findings will be developed in discussion in accordance with the three core considerations outlined above: 1) the actual usability of the various features contained within the various dashboards; 2) the intelligibility of the dashboards, the concepts underlying their organisation, and what they were designed to do; 3) the extent to which the dashboards might be easily accommodated within journalistic working practice and how far they might augment the current tools journalists are making use of on a day-to-day basis.

5.1 The Formative Evaluations

As mentioned above, a total of three formative evaluations were undertaken over the course of 2016, with various specific elements tested on a more ad hoc basis as the need arose. All of these evaluations related exclusively to the PHEME Journalist Dashboard. In this section we shall briefly present the outcomes of those various evaluations, and the implications these were seen to have for ongoing development.

5.1.1 First Formative Evaluation

It should be remembered that the version of the dashboard evaluated in the first formative evaluation was working with canned data that had already been manually annotated and other elements were at this stage hard-coded to show what they might look like.

- *Relation to Journalistic Workflow and Existing Resources:* Something that was stressed in this evaluation was the way in which time pressure can shape the nature of the work on a newsdesk, with anything that delays access to resources being potentially problematic. This had implications for keeping available data up-to-date. As a complement to other resources such as newswires and other social media platforms, this early prototype seemed to provide a good fit.
- *Provided Features:* A number of proposed features were presented to the journalists and most of these received a favourable response. These included the notion of a **map** that could display the geographical source of tweets, which could then be directly selected from the map. This uncovered the importance of localization in newsdesk work. Another positive feature was the presentation of **associated documents** and media that were linked to the original tweets, which offered the scope to rapidly bring together a range of related sources in one place. Author details were more problematic as these often drove the journalists to Twitter itself. Features that were of less obvious merit were the presentation of a **conversation history** which was viewed as more appropriate for features writing where one had the time to explore content in greater depth, and the possibility of entering **user comments** which the journalists doubted they would use.
- *Absent Functionality and Requirements:* Several important types of functionality were not available in this prototype and were considered by the journalists to be critical

requirements for future iterations. The most notable of these was the absence of any capacity to conduct a **search**, which needs to return results quickly in newsdesk work. Related to this in terms of being able to tailor the display of results was a need for **filtering** and **sorting** functionality in future versions. **Veracity** assessment in this version was seen to lack any indication of confidence in the suggested assessment or indication of how the assessment was arrived at. Representations of trends over time of both veracity scores and topic popularity were considered useful. Other requirements included the capacity to **embed** tweets directly from the dashboard into their own publication tool without having to go to Twitter, the highlighting of recognized factual content in the tweets, and the capacity to **save** and recall specific tweets or groups of tweets in the dashboard.

5.1.2 Second Formative Evaluation

The second formative evaluation was compromised by a number of technical issues, such that it was not possible for the journalists to actively use the dashboard within a typical workflow (for an indication of a typical newsdesk workflow, see D8.1 and the use cases in D8.3). The key objective of this evaluation was to present the journalists with live rather than canned data. Troubles here served to highlight critical requirements for future versions of the dashboard. These covered four principal dimensions:

- *Liveness:* Technical issues caused the refresh rate to be slow and patchy and even, on occasion, to stall. This brought to the fore the need for up-to-the-minute data when working on news which is, by definition, current.
- *Search:* Search functionality had still yet to be built in and provided keywords were either too generic or generative of clusters that seemed to be made up of largely un-related tweets.
- *Coherence:* Related to the preceding point, clustering of the tweets was making use of an algorithm that had no relationship to the kinds of reasoning the journalists would use, so results appeared to lack any obvious coherence.
- *Temporal Cohesion:* Exploration of the various tweet clusters revealed a problem in binding together conversationally related tweets that fell outside of a certain time window. Tweet relationship was noted to be central to the identification of unfolding rumours so this was an essential element to repair.

5.1.3 Third Formative Evaluation

The third formative evaluation in September 2016 sought to:

- accomplish what had not been possible in the second evaluation, i.e. to have journalists use the dashboard in the context of their normal workflow, using live data; and
- assess progress regarding the key requirements uncovered in the previous evaluations.

With regard to the latter point, in this evaluation journalists were asked to assess specific features of the dashboard in order to verify the various features so far developed. It also provided an opportunity to have a German journalist sit with a German version of the dashboard as one of the project goals was to have the dashboard working in both English and German. Findings from this evaluation served to underscore the broad effectiveness of the design objectives now adopted and

to emphasize areas where further work would be needed to ensure the dashboard was fully functional by the end of the project. Areas where it was clear the most work would be required were: time performance; the presentation of search results; clustering; the different language pipelines; and the presentation of the landing page.

- *Time Performance and Search Results*: A feature of the dashboard as it had been re-designed was that users could now set a topic running that would then be populated with tweets organised into related groups, now to be called 'Phemes'. Users entered topics as requested but encountered two key difficulties in relation to this:

    a. tweets were not returned by the system against each topic within an acceptable timeframe (indeed, it was considered to be too slow even for feature writing let alone newsdesk work);

    b. users naturally sought out topics they were interested in and entered keywords they considered suitably related to those topics, but the system did not adequately return results against those topics.

    Solutions to this that had to be adopted were searches for topics on major international news events with restricted expressions, which is not optimal for natural use. SwissInfo journalists who work in **multiple languages** also noted that the return of results was restricted to English tweets, whereas they would typically be looking for results across a range of other languages, all of which might be drawn upon in the construction of a story. An additional source of trouble here was that the system only sought to collect tweets against the topic specified as they appeared *after* the topic had been started. **Historical tweets** relating to the topic were not collected and presented. This made it clear that both default and specifiable start times for searches had to built in for future versions. Beyond all of this, the fact that the system was struggling to locate results was not always evident to the users. It was not straightforward to know whether the system had frozen, was still working, or had found everything it was going to get. **Feedback** was therefore required in future so that users could reason about system behaviour and adapt their expectations or even topic searches accordingly.

- *Clustering:* There were three separate issues identified with regard to how tweets were currently being grouped within the clusters called Phemes:

    a. One of the issues was the overall **coherence** of the grouping. Participants found the rationales for the grouping non-obvious and struggled to make sense of the groups as a consequence. A similar issue had already been identified in the second formative evaluation. A related problem was that it was not clear why the tweets within each group were being presented in the **order** they were seeing them in and whether this might be indicative of anything important or whether it was purely random.

    b. A second issue identified was the displayed **title** of the Pheme and, relatedly, the notion of a **featured tweet**. For a start the titles contained extraneous characters generated by the system and lengthy URLs etc. Just the basic text of a tweet was considered to be preferable as a minimum and, ideally, something that might more effectively summarize the content and thereby give a clue as to the rationale underlying the grouping. The notion of a featured tweet as representative of the group was also potentially problematic in that the tweet chosen as the title was

often not clearly related to other tweets within the group. Indeed, this form of presentation tended to encourage an assumption that the other tweets would somehow be responding to the title tweet, which was not the case. A central part of the problem here was that the grounds of selection for a featured tweet were not evident to users but, rather than simply see the presentation as random, they attempted to make sense of it *somehow*. If tweets are to be used as titles for a Pheme, explicitly referring to them as a sample would seem to be a safer way to proceed.

c. The third issue related to a preference to have an organisation in the clusters that captured a sense of originating tweets and the unfolding conversations or threads that these tweets generated. The term **'thread'** was also seen to be potentially problematic so it was considered preferable to avoid explicit mention of this term within the dashboard in future. Overview information about the clusters was also discussed and it was noted that a **controversiality score** would be helpful and that the provision of a **veracity score** should be given a very high priority.

- *Language Pipelines:* For the journalists working at SwissInfo, being able to see results in multiple languages and to filter by language was of particular importance. Even journalists writing news in English routinely make use of materials written in German and French. This is because much breaking news and early content in Switzerland appears first in these other languages. The fact that the dashboard was only presenting tweets in English was therefore considered to be a potential limitation.
- *Presentation of the Landing Page*: At this point the landing page for the dashboard was presenting current Phemes from Topics that were already running. When users wanted to collect new tweets relating to a new topic, they immediately sought to do so using the search box at the top of the landing page, understanding it to be a means of searching Twitter in general for that topic. However, the actual results of searches using this device were all derived from the content already displayed, which led to significant confusion. It was therefore recommended that a **new topic creation box** should appear on the landing page together with a view of what topics were currently running, rather than a presentation of current Phemes.

5.2 The Summative Evaluation

As noted above, a final summative evaluation was conducted with a total of five journalists during February and March 2017. Each of the journalists was asked to provide some demographic and background information. Note that, in view of the small sample size, many standard demographic questions would not have been meaningful. Instead we focused on obtaining information relating to their prior work experience, especially regarding their work as journalists. Here is a summary of the results.

How old are you?	25-34 (4)*	45-54 (1)	
Gender	F (3)	M (2)	

How long have you been a journalist?	Under 5 years (1)	Between 6 & 10 years (3)	21 years or more (1)
Have you always worked in news organisations?	Yes (2)	No (3)	
If the answer is no, what other kinds of organisations have you worked in?	University / Health Department of Cantonal Administration	Trade Association / Magazines	Production companies (Radio & TV) / Radio stations / Also many other jobs: shops, construction companies, restaurants, etc
Have you ever worked as a freelance journalist?	Yes (3)	No (2)	
How often do you work on the newsdesk or work shifts to produce quick stories?	Weekly (3)	Occasionally (2) (But one had previously worked on a newsdesk full-time)	
What percentage of your overall work would you say is newsdesk / quick story work?	About 50% (1)	Less than 50% (4) (but see above)	

* The number of respondents out of 5 is given in brackets

Table 1: Background Information Regarding the Evaluation Participants

This evaluation included assessment of the fact-checking dashboard, Hercule, as well as the PHEME Journalist Dashboard. As this evaluation was designed to uncover the effectiveness of the point the dashboard design had been able to reach within the project, it was **more comprehensive** than previous evaluation exercises. In particular, as well as continuing to assess the usability of the dashboard within the normal workflow, it was designed to assess each of the principal features, the experience of using the dashboard in its own right, and how well the dashboard compared to other resources journalists are used to using. Results here are also derived from both responses to **questionnaires** and to in situ observation and commentary. For the larger part the questionnaires posed questions that requested ratings on a standard 5-point Likert scale as this was considered to be a widely familiar tool and therefore not demanding too much prior explanation. Some questions demanded simple Yes or No answers. Throughout, the participants were also invited to write comments and a few of the questions required full written responses (see Appendix IV). Each of the sections in the Summative Evaluation were subject to **time limits**. If participants did not complete set tasks within the allocated time for the section, they were asked to move on to the next one and any unfinished tasks were left undone. This was taken itself to be indicative of whether the journalists were having trouble using the dashboard for the section in question. In the event this restriction only resulted in 2 journalists not completing the final set of 4 tasks in Part 1 of the evaluation for the PHEME Journalist Dashboard. In the following sections we shall be looking at the PHEME Journalist Dashboard and Hercule separately and will also present results from the questionnaires first of all and then elaborate upon them wherever relevant by drawing upon the in situ recordings.

5.2.1 The PHEME Journalist Dashboard

The PHEME Journalist Dashboard had been significantly refined since the previous, formative

evaluation and many of the issues identified in that evaluation had been repaired and new features implemented. Nonetheless, to ensure testing could proceed reliably and in a fashion that would be comparable across all of the participants, some safeguards were introduced. One of these was the use of a **pre-collected and stored dataset** relating to the Charlie Hebdo attacks in Paris in 2015 for the testing of the individual dashboard features. These tweets were streamed as if they were live data and processed in real time by the PHEME dashboard and integrated backend components.

Dispensing with the live Twitter feed meant that unpredictable variations in performance would not unduly influence the results. This also provided a means of being able to pre-specify some tasks with known outcomes so that journalist understanding could be assessed. The other important safeguard adopted was the setting up of topics for searches in advance for later workflow-related tasks, so that the system could be collecting results relating to 'live' searches whilst the journalists were engaged in other tasks. Additionally, journalists were instructed on the day regarding the choice of topics and keywords so as to avoid the selection of topics that might produce very few results or unduly clog the system. The results we present below will focus upon journalist assessment of:

- the usability of the dashboard;
- the intelligibility of dashboard concepts;
- the viability of the dashboard for their own work;
- the experience of using the dashboard and its look and feel and presentation; and
- the comparability between the dashboard and other related tools already being used.

5.2.1.1 Usability

	not at all	not very	moderately	quite	very
How easy was it to search for topics?			**1**[1]	**2**	
How easy was it to add a topic?		**1**		**2**	
How easy was it to actually explore the phemes?			**3**	**2**	
How easy was it to uncover how many phemes were in the topic?				**1**	**4**
Was the filtering easy to do?			**1**	**2**	**2**
Was the sorting easy to do?				**1**	**3**
How easy was it to locate which Pheme was most active?		**1**	**2**		**2**
How easy was it to locate which Pheme was the most controversial?		**1**	**1**	**2**	**1**
How easy was it to locate how many images were in the Pheme?					**5**
How easy was it to locate the number of verified authors in the Pheme?	**1**				**4**
How easy was it to use the keyword search feature?				**2**	**3**
How easy was it to save the results of your search?		**1**	**1**	**2**	**1**
How easy was it to get to the detailed view of the Pheme?				**2**	**3**
How easy was it to locate the veracity label?				**3**	**2**
How easy was it to locate the Linked URLs?			**1**		**4**
How easy was it to locate the images?				**1**	**4**
How easy was it to find the location of the authors?	**1**			**1**	**2**
How easy was it to open a conversation in Twitter?		**1**			**2**
Please rate how easy the dashboard was to use overall (where 0 is not at all easy and 4 is extremely easy)			**2**	**3**	
Please rate how well laid out you considered the dashboard to be			**2**	**3**	

[1] Note that all figures provided relate to the actual number of journalists giving that response. The size of the sample does not allow for meaningful percentages or statistical assessment of the results. The figures nonetheless provide a sense of the extent to which there is contiguity or spread in the views of practiced professionals who would be potential users of the dashboards.

Did you encounter any performance issues with the dashboard?	**Yes** **2**	**No** **2**

Table 2: Questionnaire results regarding usability of the PHEME Journalist Dashboard

In table 2 we can see the questionnaire-based results from the summative evaluation that related specifically to matters of usability. Unsurprisingly, features of the dashboard that simply required users to locate certain bits of information were rated as the easiest things to do. For instance, locating **how many Phemes** were present in a specific topic required no more effort than spotting the associated number in the navigation bar. Four journalists rated this task as 'very' easy and 1 as 'quite' easy. One of the journalists did, however, momentarily not see the associated figures and looked instead at the list of Phemes, saying 'Do I have to count?' All of the journalists found locating how many images were present in a Pheme 'very' easy. As one commented: "This was easy because it's written there". When it came to finding out how many **verified authors** were present in a Pheme, 4 journalists also rated this as 'very' easy. However, it should not be assumed that details listed in a small font in a list of other content in an overview are necessarily easy to locate. One of the journalists had a lot of trouble seeing this piece of information and rated the task as 'not at all' easy. All of the respondents selected Phemes where there were no verified authors. A small issue that arose here was that the Phemes merely indicated whether there were tweets from verified authors present. As one respondent commented, to know what proportion of the tweets were by verified authors "It only says yes or no… I would have to count them."

Whilst it is similarly just a feature of the overview information, the **veracity label** for a Pheme presented a little more trouble. 2 participants found this 'very' easy to do and 3 only 'quite'. The small disparity here resulted from the fact that for the first few participants the veracity label was mistakenly positioned against individual tweets rather than as part of the overview of a Pheme.

Associated content for the various Phemes was largely considered to be easy to locate. 4 participants found it 'very' easy to find both the **images** and the **linked URLs** and only 1 had more trouble, rating this task 'quite' easy for images and 'moderately' easy for the links. Despite the location of the images being viewed as largely straightforward, there were some puzzling issues here. In one case the presence of an image was reported but it wasn't actually visible, which caused the participant to speculate: "maybe they deleted it or there's a bug". In another case the overview stated that there were 17 images, but only 1 was visible. Here the participant wondered whether the picture had simply been retweeted 17 times, but it was hard to reason this through with the information available. More problematic here was the **map** providing the locations from which various authors were tweeting. Two found this task 'very' easy, 1 found it 'quite' easy, but the other 2 failed to locate any authors completely. This was because the map functionality was not working at all for one evaluation session, and for another the participant had chosen Phemes where no localization was possible. It should also be noted that on the whole we were finding that latitude/longitude information was only available for about 1.5% of the tweets streamed.

Other aspects of the usability of the system provoked more varied responses. Not all of the participants immediately saw where to click on a Pheme to be able to drill down to its more **detailed view**. Three participants found this 'very' easy but 2 rated it as 'quite', and it was visible that most of them had to try clicking in several places before they found that the correct

procedure was to click on the Pheme title. This is indicative of a need to provide clearer cues for navigating through the various levels. Using the **keyword search feature** also produced mixed responses of the same order – 3 finding it 'very' easy and 2 'quite'. The slightly larger number of positives here, however, disguises some important issues that arose with the search functionality for the first two participants. In these initial sessions, a simple keyword matching search was applied to the Pheme titles. This meant that participants were searching for terms that were present within the tweets themselves but were nonetheless getting back no results. Manual exploration of the content of the Phemes made it clear that the search was not catching everything. One participant in particular associated this with a bigger issue that we shall return to:

> "It would be nice to be able to search all of it. Especially now when the cluster is not well-defined. If I had a cluster about the travel ban it would be alright but here it's hard to drill down without it".

For later sessions, a full Boolean search across all of the tweet content was implemented and no further issues with the results arose. When it came to **saving the results of the searches**, however, there was a much wider spread of outcomes (1 'very'; 2 'quite'; 1 'moderately'; and 1 'not very'). The basic problem here was that it was not immediately evident to all participants how searches might be saved because the relevant button was not adjacent to the search box itself. It was further compounded by the menu associated with the saving of a search containing redundant material from a previous implementation of the dashboard for a different group of users.

The **filtering and sorting** functionality provided in the dashboard also received varied responses from the users. For most the actual application of the sorting categories was deemed 'very' easy (by 3), though 1 user was a little less enthusiastic, rating it only 'quite' easy, largely because it took him some time to actually see where the sorting menu was on the page. However, there were glitches here that were not reported on the formal questionnaires. For instance, it was evident that not all of the users were comfortable with the **default sorting** being by 'controversiality'. It was suggested that defaulting to 'size' would be more suitable:

> "If I was scrolling through the Phemes to start my research I would pay attention to the size – how many there are in it. I'm interested in size because it indicates significant discussion"

Users were also troubled by the selecting of alternative sort options not always taking them to the first page of results and some users found the overlap of search and filter categories confusing:

> "It was quite easy – The only thing is, in the filter windows it's easy but a little bit hard to understand what I'm filtering versus what I'm sorting"

In relation to this it was suggested that the filters might be augmented:

> "It may be cool to have a language filter as well/ And perhaps based on content: if there is a weblink or a picture. At the moment the filter and sort are much the same"

When it came to actually using the **filters**, 2 participants found this 'very' easy to do, 2 'quite' easy, and 1 'moderately'. In all cases here it took some experimentation for the filters to actually be seen to work properly. The presentation was based on sliders and the initial tendency was to slide one of these to its maximum point. However, this typically meant that users then saw no results at all. They then had to experiment with moving the slider back to a point where it might have some impact. As the point where the filters will come into play will almost certainly vary across topics, this kind of exercise would need to be repeated and was time-consuming. Another issue here was the position of the filters in proximity to the search box and the inevitable

expectation that the two were therefore somehow related. It was also an overhead to clear the filters after their use and this is something that would ideally be made much easier.

Some features demanded reasoning beyond what was simply visible in a Pheme overview and responses to these tasks were also more variable. For instance, identifying which Pheme was the **most active** demanded making an association between this and size and then sorting according to size. Two journalists found this task 'very' easy, but 2 rated it as only 'moderately' easy and 1 as 'not very'. We shall return to the issue of understanding activity in the next section. Getting to **conversations** in Twitter also involved some work. Specific tweets offered the scope to 'Open the Tweet'. However, this alone did not take users into Twitter and if they did not get to the original Twitter posting, they were unable to see the full Twitter conversation. This was compounded by the age of much of the data being used in the evaluation because many tweets had now been deleted and were no longer visible. Not all journalists attempted this task because of the time limitation. Two of those who did found this task 'very' easy, but the other one said it was 'not very' easy at all. A clear outcome of this was that journalists prefer to be able to open tweets within Twitter itself, something that was reinforced by the evaluation of Hercule where this *is* what happens.

In some cases locating the required information was rendered problematic by a lack of clarity in the responses. Pheme **controversiality**, for instance, was at first sight easy to uncover. However, multiple Phemes were returned with identical controversiality scores and users were unsure how to distinguish between them. When asked to locate which Pheme was the most controversial, the returned ratings were therefore spread across 1 'very', 2 'quite', 1 'moderately, and 1 'not very' easy.

Somewhat more surprisingly, how to actually explore Phemes themselves, which involved clicking on a Pheme title and then simply scrolling up and down the returned contents, was not rated very highly. 2 journalists here found this 'quite' easy, and 3 only 'moderately' easy. One problem here may have been the fact that not all of the journalists grasped what a Pheme was in itself straight away (see below). Some also took a while to discover that clicking on the Pheme title opened it. One further reported issue was the fact that when clicking to a different page of results the list started at the bottom rather than the top.

Another feature that was generally less highly rated was searching for a topic on the landing page and adding a topic. For the former 2 found this 'quite' easy, and 1 'moderately' easy (2 journalists did not undertake this task because they did not need to set up topics for the subsequent workflow section of the evaluation), and, for the latter, 2 found it 'quite' easy but 1 'not very'. One journalist explained part of the problem here:

> "It's quite easy to add a topic. The button is clear. But if I'm wanting to make a relationship between what I just searched for and what I'm now adding it's a little bit difficult – but once you get it it's clear"

The work involved in actually adding a topic was also potentially troublesome, with some users not realising that they had to add a description, and some finding it hard to save a topic once it had been created. On top of these issues, seeing how to select a specific topic and have it displayed was not evident to all of the users and a couple of them were confused to find that they had all the Phemes from several different topics displayed at once.

Despite the above issues, overall rating of the layout of the dashboard was not too bad, with three users rating it 3 (where the highest was 4) and 2 rating it 2. Identical responses were also given when asked to assess how it easy it was to use overall.

Participants in the evaluation were also asked to provide feedback regarding what features they felt could be presented more effectively, what features they had trouble with, and whether they had encountered any performance issues. These largely reflected the above concerns and included: starting and stopping topic searches (which was mentioned four times); sorting and filtering (also mentioned four times); time performance (mentioned three times); getting back to the landing page (mentioned twice); displaying the map (mentioned twice); and the presentation of the Pheme list and its contents (also mentioned twice). A couple also had trouble with the dashboard physically crashing and having to start over.

5.2.1.2 Intelligibility

	not at all	not very	moderately	quite	very
How easy was it to understand the Explore Social Media feature?			**2**	**1**	
How easy was it to understand the notion of exploring phemes in a topic?	**2**	**1**	**1**	**1**	
How easy was it to understand the filtering options?		**1**	**1**	**3**	
How easy was it to understand the sorting options?	**1**			**1**	**2**
How easy is it to understand what the Pheme 'Activity' means?	**1**	**1**	**1**	**1**	**1**
How easy is it to understand what Pheme controversiality is about?		**3**	**1**		**1**
How easy was it to understand the Pheme controversiality score?	**1**	**3**		**1**	
How easy was it to understand what the Pheme was about?	**2**	**2**	**1**		
How easy was it to understand who counts as a verified author?			**1**	**1**	**3**
How easy was it to understand the keyword search feature?				**2**	**3**
How easy was it to see why the threads in a pheme were grouped together?		**4**		**1**	
How easy is it to understand what a veracity label is telling you?		**3**	**1**	**2**	
How easy is it to understand the notion of Linked URLs?				**2**	**3**
Please rate how easy you found it to be to understand all of the different features			**3**	**2**	

Table 3: Questionnaire results regarding intelligibility of the PHEME Journalist Dashboard

In no instance was there complete agreement about the intelligibility of the various features within the dashboard. Nonetheless, some aspects were given relatively high ratings. An encouraging result here was the generally favourable response given to the **keyword search**

feature. Previous evaluations had all uncovered issues with the search functionality and it was considered to be a critical component to fix. Three journalists now found using this feature to be ‘very’ easy to understand, and the other 2 reported it to be ‘quite’ easy. However, some slight confusion still remained regarding just what was actually being searched.

At the **Explore Social Media** level, the search was directed at Twitter in general. However, once a topic had been opened, the search related to the contents of the associated Phemes and, in most cases, this had to be explained. It was also the case that the broader assessment of the Explore Social Media feature was less highly rated for intelligibility. One journalist rated this as ‘quite’ easy to understand and 2 others as only ‘moderately’ (2 journalists skipped this task because they worked exclusively with the canned data). This reflected an ongoing uncertainty about how to proceed when arriving at the initial **dashboard landing page**, which suggests there is more work to be done to guide users into use of the system.

Some of the information presented within the **Pheme overviews** was considered not just easy to find but also easy to understand. Linked URLs for instance were grasped without much trouble, 3 finding this ‘very’ easy to understand and 2 ‘quite’. Three journalists also found the notion of **verified author** unproblematic, reflecting their general familiarity with Twitter and how it operates. One found this ‘quite’ easy, having initially not recalled what a verified author was, and 1 journalist with less familiarity found this only ‘moderately’ easy.

Interestingly, the capacity to understand **‘Linked URLs’** did not necessarily carry over into an understanding of what URL count might mean in the sorting options. The intelligibility of the **sorting** options caused trouble in a number of cases, and this carried over into assessment of the **filtering** options as well. So, for the sorting options, 2 journalists rated the intelligibility as ‘very’, 1 ‘quite’, and 1 ‘not at all’. For the other journalist a technical issue meant that the sorting options were not available on the day. For the filtering options the ratings were 3 ‘quite’, 1 ‘moderately’, and 1 ‘not very’. We have already intimated above that there may be issues to deal with regarding the sorting and filtering options on offer and their extent to which they overlap. The finding here reinforces this point as a number of the terms provided were not self-evidently meaningful and had to be explained.

Something that provided particular trouble for understanding was the notion of Pheme ‘activity’. 1 journalist actually found this ‘very’ easy to understand, but the responses ranged across all 5 possible options down to ‘not at all’ easy to understand, with a different response from every journalist (which is indicative of confusion on its own). The intended understanding here was an association with the overall size of Pheme, which indicates how much discussion it has generated. However, in their comments had other quite reasonable possible understandings such as the average number of tweets per second or how recently a Pheme had been updated. This is clearly terminology that needs to be thought about further and clarified in its presentation.

One of our core interests in work package 8 has been providing journalists with a way of readily identifying Phemes that are particularly generative of debate and disagreement. This is reflected in the notion of controversiality and the associated controversiality score that is displayed for every Pheme. However, not all of the journalists immediately arrived at this understanding of controversiality and had to have it explained to them. This was notably reflected in their ratings for intelligibility of the concept. When asked how easy it was to understand what controversiality is about 1 replied ‘very’, 1 ‘moderately’ and 3 ‘not very’. For the controversiality score the ratings were even lower with 1 saying it was ‘quite’ easy to understand, but 3 saying ‘not very’ and 1 ‘not at all’. The cause of the trouble with the controversiality score is visible in the

comments:

> "Yeah, in a group of 1234 tweets it's nice to have it down to ten that are over 90 percent [controversial] - That helps me somewhere. But then when I go in and I see that in fact one of the 100 percent ones is not at all controversial that makes it difficult to know. It reduces my confidence in the system. If they were really all equally controversial with a positive and a negative statement then that's fine, but that would have to be the case"

> "This one is just retweeting about the vigils held in Paris. So I don't think it's very controversial at all even though it says it is"

One journalist did try to arrive at an understanding of where the ratings were coming from, for instance because the topic was an attack but the word good was in the associated text, but then they realised this didn't really seem to be a coherent reason because the assessment was based upon multiple different responses.

It was also clear that users were troubled by how the system was arriving at a specific score for controversiality. One of the causes of this was that they could see multiple results with a score of 100% without any clear distinction between them. They also struggled to understand what the difference would be between Phemes given a score of 96% and those given 100%. All of the users did, however, comment that controversiality ratings would be very useful to them 'if they worked'.

The **veracity** labels also caused trouble. One of the problems here that we have already mentioned is that they were initially positioned at the individual tweet level rather than the Pheme level, but even setting this aside there still issues with understanding. Responses regarding whether the label made sense at all were 3 saying 'yes' and 2 saying 'no'. But when it came to understanding what the label was actually telling them 2 replied that this was 'quite' easy, 1 'moderately', and 2 'not very'. A core thing to grasp here was that the system was rating veracity as an indicator of how true or false the claims in the Pheme were overall, taking the Phemes cumulatively up to the latest tweet. So this was a value that could change dynamically within the Pheme. However, users were inclined to understanding the label as referring to individual tweets and this persisted even after the label had been moved to the Pheme level:

> "So here it says false. Why does the system think it's false?"

> "It's confident that's its false but there doesn't seem to be anything false about it"

> "How can someone being sad about something be false?"

A clear message to take from all of this is that the workings of the veracity assessment and the meanings of the labels require more careful presentation within the dashboard so that users can grasp what the system is doing more intuitively.

Another problem that needs to be addressed is the presentation of the idea of a **'Pheme'**. An early task in the evaluation asked users to explore the Phemes in a topic. However, it was uniformly the case that users first of all had to have the idea of a Pheme explained to them. 1 journalist found the notion of exploring Phemes in a topic 'quite' easy to understand, 1 said it was 'moderately' easy, 1 but 1 said 'not very' and 2 said 'not at all'. here are some of the associated comments:

> "If you're coming very fresh never having seen the dashboard, it's not very obvious what a Pheme is without an explanation"

> "It needs a definition of a Pheme somewhere. And they are not visually separated enough. They need to be more offset"

> "I have to explore the Phemes? What does this mean?"
>
> "Okay how easy to understand the Phemes in a topic. I don't understand what a Pheme would be. Please tell me"
>
> "I don't really know what Phemes are"

Drilling into the Phemes revealed perhaps the biggest intelligibility issue of all. Troubles here cut at three levels. One of these related to being able to understand what the unifying idea was behind each Pheme cluster. When asked how easy it was to understand what a specific Pheme was about, 1 found it 'moderately' easy, but 2 replied 'not very' and the other 2 'not at all'. When asked specifically what the Pheme was about one replied that they were 'not sure beyond Charlie Hebdo' and another said 'I don't really understand what it's about. It's a basket of a little bit of everything".

Another issue was being able to understand the role played by a **Pheme title**, with the presentation of the title itself also proving to be problematic, with a range of extraneous characters and noise such as URLs creeping in. Each Pheme is headed with a tweet this is assessed by the system to be representative of the Pheme itself. In other words the role of the title is not really a title as such but rather a sample of content. But the comments in the evaluation made it clear that this was not evident to the journalists:

> "It's clear it's about Charlie Hebdo and in a very large sense I guess what it's about for me is the original statement – the original tweet – and reactions to that original tweet. But when I open it I don't always see direct reactions as such, I see conversations around the topic, but there's rarely someone actually responding"

In fact the comment here shows that, when presented with a tweet as a title, the journalists were inclined to make the not unreasonable assumption that the displayed tweet had generated discussion of some kind and that what they would now be seeing was the discussion. Seeing content that made no reference to the featured tweet was therefore confusing.

However, the confusion cut deeper than this because, upon inspection, users could often see no connection between the tweets that had been clustered together at all. This often left them little better off than they were working with the overall topic the search had been based upon:

> "Everything's obviously related to Charlie Hebdo, it's not like they're saying I love pandas or something, but beyond that loose relationship I don't see much more of a grouping"

Things were further compounded here by some peculiarities in the actual displayed list of tweets within a Pheme, the most confusing of which was that **repeated tweets** were counted in the overall total of tweets assigned to a Pheme, but were not displayed, meaning that a situation could arise where it would say there were 17 tweets in a Pheme but, upon opening it, the user would only find 2.

The intelligibility issues outlined above are perhaps the main part of where further work on the PHEME Journalist Dashboard would be required. Ratings of the overall intelligibility of the dashboard perhaps reflect this with 2 users scoring it '3' out of the possible 4, but 3 only giving it a '2'.

5.2.1.3 Effectiveness for Journalistic Work

	not at all	not very	moderately	quite	very
How useful would the Pheme exploration feature be to you in your work?			**1**	**4**	
How useful would being able to see Pheme 'Activity' be to you in your work?		**1**	**1**	**1**	**2**
How useful would Pheme controversiality be to you in your work?			**2**	**2**	**1**
How useful would knowing the number of images be to you in your work?		**1**		**3**	**1**
How useful would knowing the number of verified authors be to you in your work?			**3**	**2**	
How useful would the keyword search feature be to you in your work?		**1**			**4**
How useful would the detailed Pheme view be to you in your work?			**1**		**4**
How useful would a veracity label be to you in your work?			**3**	**1**	**1**
How useful would having the linked URLs be to you in your work?				**4**	**1**
How useful would having associated images be to you in your work?		**1**		**2**	**2**
How useful would having author location be to you in your work?				**2**	**2**
How useful would being able to open a conversation in Twitter be to you in your work?			**2**	**1**	
For each of the following activities please rate the potential usefulness of the dashboard:					
Newsdesk Work			**1**	**2**	**2**
Writing Feature articles			**3**	**2**	
Freelancing			**4**		
	Not at all	Not Very Well	Okay	Very Well	Perfectly
How well would using the dashboard fit into your usual workflow?		**1**	**3**	**1**	

Would you use the dashboard in your everyday work if it were available?	**Yes** **4**	**No** **1**

Table 4: Questionnaire results regarding the extent to which the PHEME Journalist Dashboard might fit with existing journalistic practice

For the larger part, the assessments provided regarding the usefulness of the dashboard for the conduct of their work were very encouraging. For the usefulness of the detailed Pheme view, 4 journalists rated this as 'very' and 1 'moderately', and the capacity to then explore the various Phemes was rated by 4 as 'quite' useful and 1 'moderately' useful. For the keyword search feature, 4 rated this as 'very' and 1 'not very'. What needs to be understood here as the background to these results is the extent to which the journalists understand the dashboard to be a means of ***scaling down*** what they would otherwise be getting if they were searching Twitter itself, with the ideal being that the results would then be specifically tailored to their current need. The latter point is crucial. Whilst they were widely appreciative of the scope for reducing the content they would need to wade through, they had misgivings about the extent to which the tool would work as a replacement for Twitter in its current form. Here is a range of quotes from the sessions:

> "The search feature in and of itself is quite useful but I use it in Twitter often and at first glance this doesn't give me anything other than what I would get in Twitter anyway. The nice thing is it's all in one place"
>
> "I usually go to Twitter and look directly - *Would you use the tool in same way?* - Probably not"
>
> "Is it useful? I think it would be if it was a minor topic. So then it would not be there straight off if I go to Twitter. But Charlie Hebdo would be there anyway"

There were also concerns about the fact that the dashboard presented them more or less immediately with a **list of content**, rather than some kind of overview:

> "I don't see the variety of different topics. At the moment it's ordered by size. It may be interesting to have a mindmap or graphical overview of the different topics and see which relates to another and which is the biggest or most discussed. At the moment it's just a list but I don't see what's behind the list"

Some of the more specific features within the dashboard provoked an immediately positive response. The **map** option, for instance, providing a visual means of localizing tweets and the sources, was very warmly received by several of the journalists:

> "This is cool. This would be a nice thing to have"

Two journalists rated this feature as 'very' useful and 2 as 'quite'. For other journalist the feature wasn't currently working. Indeed, its performance was generally patchy in the summative evaluation sessions, which was a pity as it is a feature that has excited the journalists throughout all of the evaluations.

Another specific feature that was appreciated by several of the journalists was the collation of images associated with the tweets in a Pheme. 2 journalists thought this would be 'very' useful for their work, 2 said it would be 'quite' useful and 1 said 'not very'. The latter journalist expressed very little interest in working with images generally. Being able to see *how many* images were in a Pheme, however, was not rated quite so highly (1 'very'; 3 'quite'; 1 'not very') suggesting that it is the images themselves that are seen to be of value rather than what

their presence might indicate about the Pheme itself. The presence of linked URLs was seen to be broadly useful (1 ‘very; 4 ‘quite’), and, in relation to building content out of the associated features it was also noted that journalists would like ways of being able to quickly save links and images and then embed them in their own files. The usefulness of being able to see the number of verified authors in a Pheme was not so highly rated. 2 said it would be ‘quite’ useful, and 3 said ‘moderately’. The comments here revealed that the journalists did not attach much importance to this in Twitter itself, largely just using it to check that certain people such as politicians were really themselves when tweeting. They were more interested in whether people had a specific reputation for certain kinds of expertise, regardless of the verified author label. They also make regular use of the names of news organisations as indicators of whether to look at a tweet, but these are names they already recognise without the use of a label. Not all of the journalists had time to try out the feature that enabled them to open original tweets. Those that did were not inclined to give this a high priority for their work: 1 said it was ‘quite’ useful and 2 ‘moderately’. However, some of the comments were at odds with these ratings:

> “My instinct is always to want to see the original tweet and click on the link”
>
> “If I’m interested in the tweet itself then I would want to save it so that I can find it on Twitter”

However, we have already noted that there were issues with getting the original tweets to display for the Charlie Hebdo dataset because of their age. The journalists were also concerned that the tweet was opened within the PHEME dashboard rather than taking them directly to Twitter itself. Additionally it is noteworthy that this feature was rated higher for the Hercule dashboard where the opening of the tweet was more seamless and *was* directly in Twitter.

In the section above regarding intelligibility we noted that the journalists struggled to understand a number of the features in the dashboard, including Pheme activity, veracity, and controversiality. However, this did not mean that they felt these features would not be of value in their work. Having grasped what it actually meant 2 journalists felt that a representation of Pheme activity would be ‘very’ useful to them. 1 said it would be ‘quite’ useful, 1 said ‘moderately, and just 1 said ‘not very’. For the veracity labels 1 said ‘very’ useful , 1 said ‘quite’, and 3 said ‘moderately’. This was particularly related to factual statements where there was a general view that being able to quickly see whether a factual statement was true of false would be useful. This finding is also of relevance to the Hercule dashboard where the principal focus is upon fact-checking. Something of particular note regarding the veracity labels was that when the journalists moved to a task where they were trying to actually assemble content for a story using the dashboards as though it were part of their routine workflow they made active use of the veracity labels to guide their selection of content:

> “If I see 100% veracity here I’m assuming the philli.com’s top stories are really ironworkers and her new gig. This could be an interesting little tidbit that might get me thinking about how is the rest of the world covering Charlie Hebdo. Maybe the veracity meter would help me at least be able to assume that this philli.com thing is true as part of my argument”
>
> “Here the veracity meter now gives me the confidence to click on this. But if it was 0% I would wonder if it was worth my time”

Relatedly it should also be noted that the journalists said that their principal reason for going to Twitter was to verify and gather extra information about things they had uncovered elsewhere, for instance through the newswires. Putting the provision of veracity indicators against content in this kind of context underscores its likely value.

The usefulness of a controversiality score was rated as ‘very’ by 1 journalist, ‘quite’ by 2 others,

and ‘moderately’ by the remaining 2. In their comments the journalists expressed even stronger interest in the controversiality score:

> “It could definitely be useful for me to know if it’s a subject people are talking about”
>
> “We expect from Pheme this controversiality thing, so that score is the most important – the most useful. We’re looking for things that generate debate”

It was also notable that 3 out of 4 responding journalists placed controversiality as the most useful filtering option for them in their work, with average activity being generally the least useful. For sorting, by contrast there was no particular agreement on what would be the most and least useful, suggesting that these features are reasoned about in quite distinct ways. We shall be returning to this point in the discussion.

When asked what features of the dashboard would be most useful for their work this generated a variety of responses:

> “Sort by URL count / Sort by image count (if I'm looking for pictures)”
>
> “Ability to keep a quite specific Pheme search going over a longer period of time / Ability to sort by size and controversiality of Phemes / Veracity meter (if trustworthy)”
>
> “Search & Filters”
>
> “Seeing growing rumours / Finding pictures by location / Finding tweets from locations / Seeing reactions to events”

The journalists were also asked if any features were missing as well as what features would probably be least useful. These questions were interesting because they provided a further opportunity for the journalists to principally surface the concerns they had regarding the intelligibility of certain features. Thus the notion of a Pheme itself, the clustering, and the Pheme titles cropped up several times. Interestingly 1 journalist expressed a doubt that the controversiality score would be useful, and another expressed concerns about the usefulness of veracity labels, at least as they are currently constituted.

The journalists were asked to indicate how well the dashboard would fit with the usual workflow. There was a spread of responses here. 1 replied ‘very well’, 3 replied ‘okay’, and 1 replied ‘not very well’. However, at the same time, when asked whether they would use the dashboard in their everyday work 4 out of the 5 journalists replied ‘yes’. When asked to elaborate upon these responses they said:

> “Maybe. It's useful to look for pictures and links, but it's also a lot of work”
>
> “Ability to monitor searches over time, see conversations and reactions to a topic”
>
> “Good for checking other sources than just wires / Adding additional information”

For the journalist who replied ‘no’ the response is also interesting:

> “It is most useful for breaking news I think, which isn't every day”

This view of the dashboard being primarily a tool to use for the handling of news content is also reflected in the breakdown of usefulness the journalists were asked to provide for different kinds of journalism. For newsdesk work 2 participants gave it the highest rating of ‘4’, 2 gave it a ‘3’, and 1 a ‘2’. For feature writing, by contrast, 2 rated it with a ‘3’, and 3 with a ‘2’. For freelancing work all 4 gave it just a ‘2’ (the other journalist felt unable to respond because they had never done any freelancing).

5.2.1.4 Experience

Please rate the overall look and feel of the dashboard	0	1 **1**	2 **1**	3 **3**	4
How would you describe your overall experience of using the dashboard?	Hopeless	Frustrating	Okay **5**	Very Good	Excellent

Table 5: Experience ratings for the PHEME Journalist Dashboard

Two sets of questions were specifically directed towards the quality of the experience the journalists were having when using the dashboard. One of these related to the 'look and feel' of the dashboard in use. The other asked directly what the experience was like. As can be seen in table 5, there was a spread in the responses for the look and feel, with three scoring it in the second highest category (3), then one scoring it as (2) and one as (1). For the **overall experience,** however, there was complete agreement, with all of them rating it in the middle as 'okay'. The implications of this are that there is work to be done to make the dashboard better to look at and more pleasurable to use.

5.2.1.5 Comparison with Other Tools

How expert are you in using social media?	Not at all	A bit	Moderately **2**	Very **1**	I'm an Expert **2**
How often do you use social media in your work?	Never or very rarely	Occasionally	A moderate amount **1**	A lot **1**	All the time **3**
How does the dashboard compare with any other social media tools you are using?	Nowhere near as good	Not as good **1**	About the same **3**	Better **1**	Much better
How does the dashboard compare with other tools you use more generally (such as the newswires)?			**3**	**2**	

Table 6: How well the PHEME Journalist Dashboard compares to other tools

Two separate questions asked the journalists to make a **comparison** between the PHEME Journalist Dashboard and other resources they were using in their work. This is an important metric because it is indicative of the extent to which the dashboard might actually get used. In order to gauge the responses here we asked for some additional information regarding the expertise of the users in using **social media** and the extent to which they were already using it in their work. Generally expertise was high. Two called themselves 'experts' and 1 was positioned on the borderline between 'very expert' and 'expert'. The 2 others considered themselves to be

‘moderately’ expert. Three of them said they used social media in their work ‘all the time’, 1 ‘a lot’, and 1 ‘a moderate amount’. Understanding the various platforms used is also useful. All 5 of the users were using both Twitter and Facebook in their work. For of them were also using Instagram. Three were using Reddit, and 2 YouTube and Snapchat. Just 1 was using Google+. When asked which of these they used the most, all 5 again said Twitter and Facebook. One of them said Instagram as well, and 1 said YouTube. This means that the user group being drawn on in the evaluations had a **high degree of competence** in the use of social media as an actual working tool, especially Twitter and Facebook, with Twitter being the core social media platform currently being drawn upon in the dashboard. This gives some weight to the responses provided because the journalists were generally well placed to make a judgment about how the tool compared with Twitter itself.

When asked how the dashboard compares with other social media tools being used in the work, 1 thought it was ‘better’, 3 thought it was ‘about the same’, and just 1 thought it was ‘not as good’. This is an encouraging position to be in for a prototype tool with some clear technical refinements still needing to be dealt with. he one respondent who thought the dashboard was not as good overall also commented “but there aren't any that say if rumours are true or false, that I use anyway”. When asked to compare the dashboard to other tools in use more generally as leads for stories and back-up for uncovering further information, such as the newswires, web-searches, and so on, 2 of the journalists thought the dashboard was ‘better’ and the other 3 ‘about the same’. This, too, is encouraging as it suggests that a more stable and refined version of the dashboard might very readily be adopted, as a part of the broader assembly of tools journalists will typically draw upon in their everyday working practice.

5.2.2 The Hercule Fact-Checking Dashboard

The Hercule fact-checking dashboard had been developed primarily by members of the team at Ontotext alongside of the PHEME Journalist Dashboard. Here the objective of the dashboard was to detect claims present in textual content – at present just tweets – and to use a variety of resources to assess the likely facticity of those claims. Thus its aims were somewhat more circumscribed than the PHEME Journalist dashboard.

For the purposes of the evaluation, the dashboard was running against live data, so journalists were invited to pre-enter topics in a similar fashion to the pre-entry of topics in the PHEME dashboard, so that there was a greater likelihood of there being enough content captured for the subsequent evaluation exercise. The evaluation itself was then conducted along very similar lines to the PHEME Journalist Dashboard evaluation, with the journalists being asked to look systematically at various principal features and to respond to questions regarding usability, intelligibility and usefulness on a questionnaire. Broader questions were also posed about the look and feel of the dashboard, the experience of using it, and how well it compared to other kinds of platforms. As this was the first formal evaluation of the Hercule dashboard, the questions posed were somewhat more basic than those for the PHEME dashboard and with a more formative intent because it was assumed that further design work would still be taking place. One of the journalists also attempted to make use of Hercule alongside of the PHEME dashboard in the workflow part of the evaluation so that its potential for real-world use by journalists could be better assessed. Only the last three of the five journalists undertook evaluation of Hercule as it had not been ready for hands-on use prior to that point.

5.2.2.1 Usability

How easy was it to add a topic?	not at all	not very	moderately	quite **1**	very **2**
In the Story View click on 'Related World News'. Open one of the displayed articles. Was this feature easy to use?				**Yes** **3**	**No**
In the Story View and open one of the tweets in Twitter. Was this easy to do?				**Yes** **3**	**No**
Rate how easy you think the Hercule dashboard is to use overall	0	1	2 **1**	3 **1**	4 **1**
Rate how well you think it is laid out	0	1	2 **1**	3 **1**	4 **1**

Table 7: Questionnaire results regarding usability of the Hercule Dashboard

In almost all respects, the Hercule dashboard was rated highly for **usability**. For most features where the users were asked to say whether the dashboard was easy to use, they replied 'yes'. There was a little more reserve regarding the functionality for adding a topic: 2 rated this as 'very' easy and 1 as 'quite'. The issue underlying this was the presence of some elements in the process that the users had expected to be clickable because they looked like buttons, and some extraneous parts that appeared to have no real purpose, such as a checkbox for activating a topic. The users also made active comparison between the Hercule dashboard and the PHEME dashboard they had previously been using. They were generally more impressed by the interface for Hercule.

> "The GUI looked a bit better than the one on Pheme – more graphic, more easily understandable"

For the overall rating of the effectiveness of the **layout** in Hercule, one gave it the top score of '4', one gave it a '3', and 1 a '2'. The same results were replicated for the overall ease of use.

The journalists did, however, feel that a number of features in the dashboard could be presented more effectively. Here are their various responses regarding what should be improved:

> "Clustering / Help with keyword finding"

> "Lists of topics / Concepts views"

> "Expect to be able to click on the words / icons to open something"

There were also some broader **interface** issues. The most notable of these was the fact that the landing page was not refreshing automatically because it had been assumed that people would not stay on that page very long but rather move quickly to a specific topic.

Of wider significance, however, is that when the users were asked what features of the dashboard they had the most trouble with, they replied the **clustering** and what the clustering was supposed to be doing within the various stories.

5.2.2.2 Intelligibility

How easy was it to understand the various features?	not at all	not very	moderately	quite	very
			1	**1**	**1**
Open the Landing Page for the Hercules dashboard once again and go to Topic Management. Do all of the features here make sense to you?.				**Yes** **3**	**No**
Click on 'Edit' for one of the topics and work through the available Topic Edit tabs. Is it clear to you what you can do here?.				**Yes** **3**	**No**
Now got to the dashboard Results Panel. Do all of the features of this page make sense to you?.				**Yes** **3**	**No**
Click on View Topic for one of the topics. Do all of the features here make sense?.				**Yes** **1**	**No** **2**
Does the idea of checkworthiness make sense to you?				**Yes** **1**	**No** **2**
Now view one of the displayed stories in detail. Do all of the features make sense?.				**Yes**	**No** **3**
Now click on one of the concepts present in the story and explore the page this takes you to. Do all of the features make sense?.				**Yes**	**No** **3**

Table 8: Questionnaire results regarding intelligibility of the Hercule Dashboard

Overall the **intelligibility** of the various parts of the dashboard was considered unproblematic. The results panel, the topic management page and the topic editing tabs were all comprehensively seen to make sense, excepting the minor issues already mentioned above. However, the rating of the dashboard for its intelligibility overall reveals the presence of some more important issues. The respective rates for intelligibility here were 1 'very', 1 'quite', and 1 'moderately'. These mixed responses refer to the trouble some of them had with the **topic view**, which is accessed by viewing an individual topic. Here 2 of the users found the features did not make sense. One issue was the notion of '**Checkworthiness**', another was the various sorting options, including opaque terms such as SDQ. In fact, when probed, most users did actually grasp the basic point of checkworthiness being an indicator as to how useful it might be for them to explore a particular story, so this is probably mostly just a matter of finding the right terminology. However, the users were unsure as to the grounds for them seeing the various **story columns** on the topic page:

> "Here I don't get what I see. It doesn't make sense. I created a topic, but now I see two columns. Were they hashtags?"

> "I'm not sure what the difference is between stories and topics. I suppose there's maybe multiple stories that can go in here or something"

Drilling down into individual stories to try and make sense of them caused further trouble. All three users here said that the **detailed display** of a story did not make sense to them. One of the key issues here was the story clustering, with it not being at all evident as to why specific things had been placed within specific stories, very much echoing the clustering issues that were present also in the PHEME Journalist Dashboard. The following comments reveal the extent of confusion the clustering provoked:

> "I don't get the difference. Here there's EU parliament. Here it's the same. And in both there are ISIS tweets. I don't see difference between the two clusters. One is bigger than the other, but both are not really separate"

> "I clicked on view topic and now I've got this. So it seems like there's two columns and you'd think they'd relate to the keywords like this and maybe 'breaking retweet' is a set of something that comes with it. But it's not in here, in the keywords. But EU and ISIS are, but not together. So it may be the EU embargo or something"

Generally the participants had trouble getting through to the concept view on the dashboard, which allowed them to see a richer body of information relating to various concepts that were present within the tweets. In all cases they had to be talked through the process and were not clear on why initially clicking to the concept view presented them with a pie chart, or why it was necessary to click on labels within the pie chart to get through to the richer view. They also expressed uncertainty as to whether what they were getting in this view offered them significantly more than they might get through using conventional web searches.

A key challenge confronting Hercule is to make it clear just what the dashboard might be able to do for people like journalists when they encounter it. At first sight they do not get 'the big idea'. All three journalists had comments to make on this point:

> "It's very easy to understand but I don't really know what it does now"

> "It's easy to use but I don't get what it does help me for. I can see there's filtering but I can do that with other tools. And the clustering didn't work for our topic. Maybe the topic was too complicated or maybe we didn't use the right keywords. But it is easy to use and well laid out"

> "I'm wondering what I'd do with it. You can insert a topic and it tells you more about it. Maybe you can click on it and see something else"

What this reveals is that it had somehow slipped past all of the participants that the dashboard is primarily geared towards the fact-checking of claims present within tweets. This is clearly something that needs to be made much more evident on the landing page and in how the results are presented.

5.2.2.3 Effectiveness for the Journalistic Work

Do you think the notion of checkworthiness will be useful to you in your work?	**Yes** **2**	**No** **1**
Do you think the concept view will be useful to you in your work?	**Yes** **2**	**No** **1**
Do you think the related news option will be useful to you in your work?	**Yes** **2**	**No** **1**
Do you think being able to open the tweets in Twitter will be useful to you in your work?	**Yes** **3**	**No**

How well would using the dashboard fit into your usual workflow?	Not at all	Not Very Well	Okay **1**	Very Well **1**	Perfectly
Would you use the dashboard in your everyday work if it were available?				**Yes** **3**	**No**

Table 9: Questionnaire results regarding the extent to which the Hercule Dashboard might fit with existing journalistic practice

As with the PHEME Journalist Dashboard there were mixed responses regarding the usefulness of the Hercule dashboard for journalistic work, some of which were encouraging, some less so. Specific features, such as being able to open tweets directly in Twitter, were viewed as being unproblematically useful. A number of the core embedded features were also fairly well received. The notion of c**heckworthiness**, for instance, once explained, was considered to be a useful thing to have by two of the three journalists:

> "Checkworthy stuff is interesting for journalists"

> "Yes… once I know what it is based on"

Being able to see **related world news** was given a similarly favourable reception, with it being particularly noted that this might provide "more sources than a regular search on internet".

As we saw above, however, the **detailed story display** provoked more variable responses, with one journalist thinking it would be useful, one thinking it wouldn't and one somewhere in-between. Clicking through to the concept view was also felt to be "maybe too much information… But some would be good".

Whilst one journalist did not express a view regarding the overall fit of the dashboard with their existing workflow, the other two responded with 'very well', and 'okay'. More importantly, all three journalists said they would use the dashboard in their everyday work if it were available. Some caveats were, however, applied to this:

> "I think if there were those features I mentioned [better overviews] it would be useful but I don't see a big part at the moment that would help in our daily business"

> "At the moment we don't have a specific tool for this. So every tool is better than nothing. But I don't have a comparison. I currently use searches… Tweetdeck… Searches on twitter. But we have no auto-monitoring tool. But if there was a monitoring tool or everyone was working the same way it would be more objective and transparent"

> "Yes we would use it if it was available – but then the clustering should be improved"

Outside of the clustering some other improvements were also suggested. One such feature was an indication of the sources actually drawn upon to provide the composite information visible in the **concept view**. As one user commented: 'journalists care about sources". Indeed there were general reservations about the current concept view, suggesting it needs further development. A more pointed and interesting observation was the following:

> "But I would like the tool to learn which keywords are important for a topic. Or have a way it can show us other keywords that are linked to the topic. Otherwise you still have to do it yourself. Filling in keywords is quite complicated in the daily work"

What this makes clear is that something that can detect and then actively propose viable

keywords when searching on a particular topic could provide a significant value-add for these kinds of tools.

For potential enrichment of the concept view the journalists proposed a few additional sources. These included national press archives, individual news organisation archives, and also news agencies.

5.2.2.4 Experience

How would you describe your overall experience of using the Hercule dashboard?	Hopeless	Frustrating	Okay	Very Good	Excellent
			1	**2**	

Table 10: Experience ratings for the Hercule Dashboard

Generally, the users of Hercule were very positive about its overall appearance and look and feel, and the experience of using it was rated 'very good' by two of them and 'okay', by the other. In particular, they felt it compared favourably on this score with the PHEME Journalist Dashboard.

5.2.2.5 Comparison with Other Tools

How does the dashboard compare with other tools you currently use?	Nowhere near as good	Not as good	About the same	Better	Much better
			2	**1**	

Table 11: How well the Hercule Dashboard compares to other tools

With regard to how the Hercule Dashboard might compare to other tools, we have already observed some pertinent points in the discussion of its usefulness. One journalist rated Hercule as 'better' than other tools, the other two rated it as being 'about the same'. But both Hercule and the PHEME Journalist Dashboard offer certain kinds of functionality that are not currently available in other tools. As journalists typically like to run a suite of tools, each addressed to different things, they are quite happy to embed these tools in that broader assembly. What is less clear is whether they would look to the dashboards as *replacements* for other tools. One, for instance, commented:

> "At the moment we are working with different platforms for that. We are not planning to have it all in one place. In fact, I'm not sure it's needed"

5.3 Discussion

In this section we look back over the whole of the evaluation exercise that took place within PHEME, starting in January 2016, and reflect upon how the dashboard has progressed and the nature of the challenges that still remain should further development be pursued in other contexts. Our discussion, whilst informed by the above concerns regarding usability, intelligibility and fit to practice, will be shaped around a set of abiding issues that have been present in one form or another throughout the course of the technical development of the dashboard. These will include:

- The desirable features to capture in a dashboard for journalists
- Relatedly, what *content* to capture (e.g. in terms of social media more broadly, associated links and images, the temporal reach, conversational origins and relationships, authorship, language, and so on)
- Timeliness, liveness and responsiveness
- Fit to workflow
- Fit to existing resources
- The relationship to Twitter
- Search functionality
- Filtering and sorting functionality
- Veracity assessment
- Representations of controversiality
- The overall look and feel and the experience of using the dashboard
- The notion of a Pheme and what to call each cluster
- The clustering itself
- Contextual use and the uncovering of the right kind of information in the right quantity at the right time

5.3.1 Desirable features and desirable content

The original organisation of the PHEME Journalist Dashboard, and the features included in its design, was developed out of requirements uncovered in the original in-depth ethnographic study of journalistic work (see D8.1). Many of these have been hung on to and the exact nature of the requirement has been further elaborated through the evaluation process. At the end of our current journey, then, we are in a reasonably strong position to be able to comment upon what the desirable features to capture in a dashboard for journalists might be.

A good deal of this is shaped around what kinds of *content* should be captured. Both the PHEME dashboard and the Hercule dashboard are premised upon the capture of **social media** and the evaluation process has made it clear repeatedly that providing windows into the vast bulk of user-generated content available through social media is desirable because it offers to reduce the overhead involved in finding both story leads and useful information. Journalists already spend a lot of time on social media and tools like this might make that task a little more tractable.

However, it does not end there. Another feature present in the PHEME dashboard from the beginning that has repeatedly been seen to be of value is the scope the tool might provide for uncovering other material beyond just social media, such as **images**, **videos** and **associated news reports**. Journalists make regular use of these things as well, but dashboards like PHEME hold out the prospect of providing a way of quickly navigating to these resources without the overhead of multiple other searches using other kinds of tools and platforms.

An interesting possibility that was only touched upon in only the most rudimentary fashion in PHEME is scoping the **temporal reach** of returned content and analysis, depending on what you are trying to accomplish. To put this crudely, when you're interested in breaking news, anything that's hours old will probably not be useful. However, if you're wanting to scope the whole trajectory of a rumour from its first appearance to the point where it has been definitively determined to be a matter of fact or entirely false, you may want to go back into material that is many months old, perhaps even older. Being able to set limits on what temporal span you might want applied to content gives you the possibility of meeting both of these ends without having to

sift through undue ‘noise’.

An aspect of content, and in particular social media content, which has proved more than most others to illustrate the variability of journalistic work is the notion of a **conversation**. Often there are originating posts and responding posts that may in their own right then generate other responses. This is captured at some level in Twitter as a conversation, which can be displayed by choice within the Twitter interface. Facebook comments allow for a similar kind of interchange and interaction of this kind has become core to social media. Elsewhere (Tolmie et al, 2015; Zubiaga et al, 2016) we have emphasized the importance of understanding these interchanges for identifying rumourous content and the dynamics through which it spreads. In the evaluations journalists have shown a mixed interest in being able to view conversations on Twitter. In the initial formative evaluation, the journalists expressed the view that seeing conversation histories was better suited to the writing of feature articles than newsdesk work because it cost time to drill down into the interactions in this way and time is something one doesn’t have when dealing with the news. At the same time, however, the journalists have been frustrated in the other evaluations, including the summative evaluation, by the absence of a clear sense of an originating tweet and subsequent tweet relationships within the Pheme clusters. Clearly there is a place for presenting conversations within the dashboard, and it is crucial for the effective tracking of rumours, but an ongoing challenge is to know just where and how to present this kind of content. At the very least it needs to be understood that the display of tweets in conversational form has to be an option rather than a default or it may force journalists to wade through unwanted content.

Authorship has also revealed variable interest. In the summative evaluation journalists recognised the possible value of having an indication of verified authors, but they did not position this as a feature of core concern. In other evaluations, however, the journalists emphasized how they would often structure their searches along the lines of looking first to content from recognized sources they could trust because this lowered the overhead of verification. Again, then, authorship matters, but its position within a dashboard display is going to need to be structured in such a way as to allow for case-by-case judgments about its possible relevance. Twitter verification is also probably not the most useful of resources. In the summative evaluation, one journalist pointed out that they were often using other criteria to judge whether an author might or might not be of interest. In the workflow stage of the evaluation, we also saw journalists actively reasoning about whether to examine tweets more closely or not based upon whether they had come from sources like the BBC or the New York Times or, on the other hand, from sources such as Fox News.

Throughout the evaluation cycle we received regular reminders from the journalists that part of how content delivery needs to be shaped from their point of view is via **language**. It is undoubtedly the case that this was somewhat shaped by the specific nature of the evaluation panel. Swiss journalists are used to pulling upon resources in German, French, and English, even if their output is often going to be only in English. Indeed, much Swiss news appears first in German and the preferences of journalists reflect this in how they set up the filters on the newswires. It is therefore quite natural that they will want to have mixed language responses available from social media. Had the evaluations been conducted in purely English-speaking countries, the importance of this requirement might not have come to the fore but it is clearly the case that a properly developed dashboard would need to honour it to be a fully effective resource. Unfortunately the closest we came to evaluating this aspect of the dashboard was the use of a German interface for part of the third formative evaluation, though the intention to

integrate a fully-functioning German pipeline had been there throughout. This therefore remains a task to be done should the dashboard be developed further in the future.

One other feature that provoked particular interest amongst the journalists, right from the first evaluation, was the possibility of **localizing** where tweet content was coming from and where tweet authors were based. The first evaluation presented this functionality in **map** form (purely as a concept rather than as a functioning prototype) and this generated immediate enthusiasm. Map features were not integrated for the second and third evaluations but a functioning map was present in the summative evaluation and it was again met with keen interest, despite a range of glitches that prevented it from working all of the time. Journalists particularly liked the capacity to see a range of information attached to specific pins on the map (author and the associated tweet as a minimum) and to be able to zoom in so that they could identify more precisely where the author was based. The underlying interest in all of this is that journalists use information about author location to be able to assess how proximate the authors are to unfolding events and their prospective status as witnesses. This is particularly useful in situations of breaking news.

Whilst the journalism use case indicates a range of content-based requirements, it should be noted that things actually go much further than this. There is also a keen demand for a variety of *analytic* resources. These are resources that do not just re-present or reorganise content from social media in some way but rather extract data from social media and then perform other operations to say something about the content. In the context of the PHEME Journalist Dashboard, the algorithms giving **controversiality** and **veracity** ratings are very much of this order. There are ways in which the map functionality might also be seen to have this character, especially if it begins to draw on some kind of named-entity extraction. The fact-checking algorithms and concept extraction in Hercule are also resources of this kind. A challenge with all of these kinds of resources is finding the right way to present them so that users can understand them, engage with them, and find them useful. Nonetheless it was visible in the evaluations that there is a strong interest in having this kind of functionality in a dashboard and the summative evaluation in particular would seem to suggest that efforts in this direction have not yet gone far enough. Hence the desire one journalist expressed to have something more than just a ‘list’. **Graphical overviews** were mentioned a number of times throughout the evaluation cycle. These might be used to capture trends over time, potential relationships, comparative measures and so on. Some of this is, of course, present in existing Pheme overviews, but it is largely in textual form and at the individual Pheme level.

Analytic resources can also service a number of ends. Clearly there are ways in which they can support verification and much of the PHEME Journalist Dashboard has been designed with this end in mind, as has Hercule. However, journalists may want to use them for more than this. In the context of supporting story development, overviews may provide the scope to ‘notice’ certain phenomena. In other words they may provide inspiration. They may also support cohesion by bringing salient points of interest together in a single place so that everything important can be seen at a glance. Analytic resources may be used to service the location of specific features of relevance. Again both the PHEME and the Hercule dashboards already do this to some extent but it has also been noted that this is one of the areas where the most work is needed. Other things that may also be serviced are: the **impact** of certain features – re-tweeting and geographical diffusion would be examples here; **currency** - for instance showing at a glance which topics are currently rising or falling in popularity (which is already a feature of a number of tools designed for journalists). Resources tracking associations (as one journalist actually mentioned during the evaluation exercise) may also serve to support the location of related

points of interest and background material.

Overall the evaluations have revealed that the PHEME dashboard has made significant progress with regard to the capture and display of content. Where it most needs development now is in terms of how it extracts and presents data from that content to give journalists analytic insights that they might not otherwise get. Hercule is further along this road than the PHEME dashboard and this was part of what was appreciated about its look and feel, but even in Hercule further development along these lines is probably required.

5.3.2 Timeliness and the fit to workflow

One of the most abiding concerns regarding the operational viability of the dashboard has been its capacity to handle **live content** in a timely fashion. Both our own studies and others (Diakopoulos et al, 2012; Tolmie et al, 2017) have emphasized the extent to which journalists may be working under time pressure when dealing with news, especially breaking news, and how this shapes all of the other aspects of their work. In particular it places stress on the capacity of systems to deliver material quickly enough for it to be of use. In the very first formative evaluation, journalists indicated that a processing time of ten minutes would not be acceptable and that, if things took this long, they would turn to other resources instead. Much of the technical work on the overall pipeline has been focused upon improving the responsiveness of the system as much as possible. However, it was notable that even in the summative evaluation any sense of slowness on the part of the system was noted by the journalists and commented upon. Dashboards such as the PHEME one and Hercule are attempting to deliver relatively sophisticated resources and to engage in challenging under-the-hood analysis, which carries implications regarding how fast a system can react. However, to support newsdesk work this still has to somehow deliver material in a sufficiently seamless fashion that journalists do not feel they are waiting on the system. The time performance of the PHEME dashboard has improved enormously, especially in the final months before the summative evaluation, but this is not a cause for complacency. Dashboards operating in newsdesk environments will always need to respect the need to deliver results as quickly as possible.

A related concern present throughout the evaluation is the scope journalists have for understanding what the system is doing. Early on it was hard for them to judge whether the system was actually working or had simply frozen. For the summative evaluation a play, pause and working metaphor was adopted, with associated icons being placed adjacent to topics. However, it was clear that many of the journalists found this kind of representation either obscure or confusing so this is an area that requires further work.

Against the above considerations, one needs to consider the fit of the dashboard to the journalistic **workflow**. As we articulated in D8.1, journalist workflows are complex affairs with a very wide range of contingencies built-in. This being the case it would be wrong to assume that dashboards such as PHEME are going to have to fit with any one single workflow. In Deliverable 8.3 we articulate three basic use cases for journalism, which are hunting for stories for the news, locating information to support news stories already being written, and feature writing. We also noted how each of these use cases results in a different workflow with different requirements potentially needing to be serviced. In particular the time pressures are significantly different for each use case, which impacts upon the range of resources journalists can make use of. Right from the first evaluation it was noted that whilst drilling down into rich conversation histories might be of interest for feature writing, the luxury to do this in-depth work is not often

present when people are working on the newsdesk. In all of the evaluations, the journalists were asked to try and make use of the dashboard as though they were actually engaged in writing a story so that we could assess its fit to their workflows. Only in the second evaluation were they actually thwarted in their attempts to do this and, for the larger part, in all of the other evaluations ways in which the dashboard might provide effective support were uncovered. The biggest challenges to doing this were: matters of performance, i.e. simply getting no returns on topics being searched for and therefore having to give up; and the coherence of the returned results, which tended to obstruct the effective assembly of relevant materials.

5.3.3 The fit with other resources

Central to the questions of workflow discussed above and a recurrent point of interest throughout the evaluations is the extent to which the dashboard is complementary to other resources that journalists are drawing upon on a regular basis. For the larger part the dashboard has scored well on this front. Right from the start journalists found it easy to work with the dashboard alongside other resources, such as the newswires, web-based searches, and Microsoft Word,. There are, nonetheless, some important considerations here.

One key point is the way journalists are inclined to orient to the dashboard within their workflow. Time and again we saw the journalists turn first to the newswires to identify possible stories to work upon. The dashboard was then looked to as one possible source for providing further information about possible leads. In the summative evaluation the journalists were somewhat sceptical about the Explore Social Media function on the landing page and rated this less favourably than some other features, which is perhaps not surprising if it does not fit entirely with their expectation of how they will use the tool. The journalists also thought that the newswires would continue to be their first port of call. This would imply that organising dashboards as a vehicle for simply browsing social media is not a high priority as this is not the way they will most often seek to use the tool. However, it is also the case that at present the dashboard does not provide overarching snapshots of analysis, such as trending stories, growing rumours, or hotbeds of controversy. All of these features might encourage a different kind of use and we have already mentioned that one journalist expressed dissatisfaction with the way the PHEME dashboard currently presents them primarily with a list.

When the journalists were asked to compare the tool with other resources in the summative evaluation, the ratings were encouraging in that it seemed likely that, given the opportunity to do so, they would make use of the tool. However, it was also clear that the tool was only being seen as a potential addition to their existing suite of tools, not as something that could be a game-changer in its own right. This also surfaced the additional interesting question of whether designers of tools such as the PHEME dashboard or Hercule can reasonably expect them to do more than this. One journalist was not sure there was any benefit to be had in moving away from the current model of using a wide variety of tools across various platforms.

Something important to consider also in this regard is the fact that both the PHEME dashboard and Hercule currently centre their attention upon Twitter. The summative evaluation confirmed the strength of place Twitter has as a preferred social media platform in the world of journalism. However, there were others mentioned. The goal in the final phase of the PHEME dashboard was to incorporate Reddit into the dashboard’s offering. Early tests of its compatibility with the dashboard appear to be very promising but it was not available in time for the evaluations. A crucial detail to be uncovered in future research is the extent to which dashboards that bridge

social media platforms may make a difference not only in terms of what benefits they may provide but also in terms of how they might be oriented to as resources within the workflow.

An associated concern that came up in almost all of the evaluations was being able to actively make use of the content of social media being presented in the dashboard without being obliged to go to the social media platform itself to extract it. Existing export functions within the PHEME dashboard were considered too limited in the kind of output they provided so there is definite scope here for further development. Additionally, it should be noted that journalists make use of a wide variety of publication tools once their stories are ready for publishing and the extent to which the dashboards might integrate with these kinds of tools has not been the subject of evaluation.

5.3.4 Searching, filtering and sorting the content

In the first formative evaluation it became very clear that search functionality was absolutely central to the viability of the dashboard as a working tool for journalists. The inadequacies of the search offering continued to be a point of concern in the other formative evaluation as well. However, by the time of the summative evaluation the search functionality was operating well and was duly appreciated by the journalists in their rating of the dashboard. This may be counted as one of the definite successes in this work package, though it should be noted that some issues do remain. The capacity to search across *all* of the content was not properly supported until the last few evaluation sessions and it was clear that this mattered. There is also an ongoing issue regarding the extent to which the dashboard makes visible just what is being searched when the search box is being used.

Filtering and sorting were also strong requirements from the outset and these have followed a similar trajectory to the search option. However, the summative evaluation revealed several points of concern to be addressed here. One issue was the legibility of the terms being used. For both dashboards, sort options were provided that were not a ready part of how journalists themselves describe the content and this needs to be addressed. The usefulness of the various options was also brought into question and some clearly matter more than others. At the same time there were noted oversights, such as being able to sort by language. Perhaps the most significant point of concern, however, is the extent to which the design currently promotes the view that both filtering and sorting are much the same thing, with many of the categories provided being identical. At heart this reflects a view that both of them are simply about organising what kinds of things you are seeing within your results. However, several journalists seemed to challenge this view and to have a sense of filtering in particular being something that is more closely allied with what they are doing when they are *searching* for content. In this respect it should also be noted that, for the newswires, journalists operate within the standard application of a variety of filters that serve to shape what content they get to see in the first place. This is often closely allied with organisational constraints about what might be appropriate topics for them to be addressing. It is also the case that these filters are constructed around the presence of key terms of interest, rather than categories per se. In other words filters are used to limit returns so that they contain things that they want to see. A broad categorical approach to filtering may therefore not be the most appropriate way to proceed.

At the end of the day searching, filtering and sorting are bound up with structuring the returned content so that they get to see according to their current need. In the evaluations this was of varied success, with the journalists often still needing to scroll extensively through returned

tweets to identify things that might be useful. It also appeared to be the case that some of the other structuring devices, such as organising the content into clusters, could work against this goal, with desirable content being spread across different containers that had to each be examined separately. This appeared to be the case in Hercule as well.

5.3.5 Veracity and controversiality

Notions of veracity and controversiality are at the heart of the PHEME project and functionality on both these scores was an important focus of dashboard design. It was also clear that the journalists involved in the evaluations understood this to be key part of what the dashboard had to offer. Actual performance on these two areas over the course of the evaluations has been patchy.

Veracity assessments in the first evaluation were seen to be useful, but there was a need to provide measures of confidence in the label being assigned and journalists had concerns that the way in which veracity was being arrived at was opaque to them. In the second evaluation no particular issues regarding veracity and controversiality were surfaced. In the third evaluation there was concern that these features were not being made sufficiently visible within the presentation of the results. In the summative evaluation it became clear that issues remain. The usefulness of these features was never brought into question and, indeed, we saw journalists actively making use of veracity scores and controversiality labels when engaging in the workflow tasks within the evaluation. It was evident that they provide an effective resource for identifying what content the journalists might want to examine in greater detail. Problems here are manifest across three levels, each of which are interlinked, and each of which suggest points of focus for future design.

First of all, and in a sense at the most superficial level, the **representation** of veracity and controversiality has to be right. Design here has oscillated between the use of numerical measures and qualitative labels. Questions have been posed about whether simple numbers or percentages work best. Discussion has also arisen regarding whether various kinds of graphical presentation might be more effective. There are pros and cons to all of these things but it was apparent in the summative evaluation that there is still work to be done here and it would seem likely that this is something that requires independent testing as a topic in its own right.

The second area of concern is the **intelligibility** of the results the system provides. Obviously a part of this is wrapped up with what the results look like and how they are expressed. However, there is a more fundamental point here. The results have to *make sense*. That is, users need to be able to look at the results and see how they might have been arrived at. They need to be accountable. It was notable in the summative evaluation that in many instances users were confounded by the labels and scores given and could not unravel how they related to the content they were seeing. It should be noted here that it would not seem to be the case that users expect the system to be perfect on this score. They understand its potential role as a guide or an indicator. But if they cannot grasp how the system is arriving at its assessments, it is hard for them to know when and where they might need to disregard it.

The third area of concern is the degree of **confidence** users might invest in the system. This is not directly to do with the confidence measures the system may provide for the results it is itself delivering, though this may play a part in it. If the system is seen to be producing results in a predictable, consistent, and coherent way, then users will be more likely to trust it. This goes for all features, including the labels or scores and the given confidence measures. An important part

of the trouble users had with both verification and controversiality was that there was a notable lack of consistency in how the system was behaving *from their own point of view*, with apparently similar things being given widely divergent assessments. The point about it being consistent from their own point of view is important. There may well be consistent technical reasons for the system to be behaving in the way it does. But they do not have the technical reasoning available to them, as was noted in the preceding point. Instead the system has to appear to be consistent from a common-sense point of view, which is altogether more challenging. Common sense of course can incorporate many kinds of reasoning, including quite specialised ones. If the cohort of users of the system see it as regularly producing certain kinds of results in the same kinds of ways according to procedures they have been made aware of, then they will be in a position to have a shared body of reasoning about its behaviour so that they can say, 'oh it's just doing that because of this feature or that feature'. This is enough for them to be able to trust the system enough to use it even if they do not always invest *belief* in its results. So the critical focus of design really needs to be upon making the system accountable, as was discussed in the second point above.

The point here about being able to trust a system even if you dismiss its result is also important. The interest journalists have shown in verification and controversiality in the project is ultimately about being able to arrive at a manageable body of results that might be open to further exploration. The system is there to assist them in throwing away the dross so that they can look more closely at what is left (this point is returned to in section 5.3.8). For the final judgment about veracity they will in all likelihood continue to make use of a variety of verification techniques, including ones they already draw upon, such as who is saying it, how many trusted sources are saying it, how many first-hand witnesses are saying it, what features are independently open to being confirmed, the accuracy of specific figures and claims, and so on. It is not likely that these final judgments will ever be ceded to automated systems entirely.

5.3.6 Look and feel and experience

As with other aspects of the dashboard, the evaluations presented a variety of responses to the overall look and feel of the PHEME Journalist Dashboard and the experience of using it. The responses to the look of the dashboard were largely positive in the first formative evaluation, though there were noted layout errors. However, in the second formative evaluation a more stripped down interface, with keywords in boxes attached to each cluster, left the journalists far less impressed, despite explanations that this was an interim prototype exclusively designed to test their interaction with live content. A range of important concerns uncovered in the second evaluation meant that the third version the journalists were presented with was altogether more polished, though there were still concerns about some elements of the layout and ordering of pages. The version the journalists used in the final summative evaluation was relatively close to that used in the third formative evaluation from an interface point of view, though certain elements had been improved and other functionality added in. For the larger part the journalists were satisfied with this presentation, if not excited by it.

Evaluation of the Hercule dashboard provoked comparison between the two dashboards and it was clear that users preferred the look and feel of the Hercule interface and the presence of a richer array of graphics. As we noted above, one journalist was disappointed that the PHEME dashboard largely presented results as a list. There is a clear need to make the PHEME dashboard more inviting and to present the results in a more varied and graphical way. However,

it should be noted that some parts of the dashboard design were equally criticised across both dashboards. The story clustering in Hercule, for instance, was not received any more favourably than the clustering in the PHEME dashboard.

5.3.7 Titling and clustering

Across both dashboards the feature that perhaps provoked the greatest number of usability issues was the clustering of tweets into groups and the associated titling. Some of the push towards trying to present more coherent groupings of tweets came from the first formative evaluation, where journalists struggled with what they found to be a confusing distinction between news 'events' and the sub-grouping of threads of discussion which were termed 'stories'. Re-worked clusters for the second formative evaluation unfortunately fared no better and, if anything, resulted in even greater degrees of confusion, with users unable to make sense of what unified any particular cluster. In the third formative evaluation the idea of presenting groups of related tweets as Phemes had started to take shape, and by the summative evaluation this had become one of the core aspects of the presentation. However, in both the third formative evaluation and the summative evaluation users continued to have trouble understanding both the notion of a Pheme and the grouping within it. It would appear then that this remains an open issue with regard to the dashboard. It should also be noted that the clustering presented similar issues for the Hercule dashboard as well.

An associated issue that was again present in the evaluations for both the PHEME Journalist Dashboard and Hercule was the naming of each cluster. It was felt that the easiest strategy here was to use an algorithm to identify the tweet that seemed most representative of a cluster or that was most significant within the cluster in some way. This was then presented as the Pheme title. We have already seen in our discussion of intelligibility in Section 5.1.1.2 that this caused a number of problems. One of these was the difficulty often of finding the relationship between the title tweet and the other tweets within. Another was the expectation that it provoked that all of the other tweets would somehow be responding to it. A further problem that persisted here was the presence of noise, with various unwanted character strings, web-links, mentions, and hashtags appearing in the title and making it harder to read the text. However, it was notable that in Hercule, where the title text was tidier, there were still issues with understanding how the story title related to the contents within.

In some ways this is surprising because from a technical point of view there has been significant progress in the performance of the clustering algorithm being used in PHEME (Derczynski et al, 2015). However, here we have a problem that goes to the heart of the matter when it comes to trying to design dashboards for situated use. Both the system and the user are attempting to achieve similar ends. They are trying to cut through the need to wade through excess information by getting it bagged up in appropriate ways. For a system to accomplish this, it has to work with a set of formalised rules or procedures whereby it can structure the content in a form that is calculated to be meaningful. There are flexible ways of going about this that are able to allow for a range of possible contingencies, but ultimately the system has to anticipate what users will see as meaningfully going together. The trouble here is that there is an assumption that the technical reasoning built into the system can somehow stand as a proxy for the social reasoning that users are engaged in when handling situated tasks. This is by no means always the case. Sometimes it works, but more often it doesn't. There is a significant research challenge built into this that PHEME, as a project, has already contributed to in important ways. But there is much work that

remains to be done. Efforts are already being undertaken here to try and explore this issue further by examining in detail the reasoning journalists themselves adopt when grouping different kinds of material together. The hope is that it will be feasible to design more adaptive systems that can use various techniques, such as machine learning, to arrive at something that more closely matches how people themselves approach the clustering of content.

5.3.8 Servicing the situation of use

The troubles presented by clustering that were discussed in the previous section actually help to bring into view a problem confronting all dashboards seeking to support journalistic work that resonates across all of the issues outlined above.

At core journalists understand dashboards such as PHEME and Hercule to be providing them with a means of managing the vast quantity of potential material available and bringing it down to a manageable size. The content of the bit that is brought down to size matters. It has to be the right kind of information, it has to be presented in the right quantity, and it has to be made available at the right time. A quotation from one of the journalists trying to use the PHEME dashboard to actually construct a story encapsulates this concern:

> "Right so. I guess now with only 15 results I would probably just scan them to see if anything looks relevant"

Scrolling through reams and reams of tweets to find the ones that might be useful to you is laborious. So you want the system to deliver just enough that actually inspecting them might seem worthwhile. But they can't be just any 15 tweets. They have to be 15 tweets that are related to the specific task you are engaged in now.

All of this may at first sight be relatively obvious but, as our discussion of the clustering problem illustrates, it actually has wrapped within it an enormous challenge for any system to overcome. One of the key problems confronting designers of dashboard type tools for journalists is that that journalistic work is actually enormously varied and shaped around a wide variety of different contingent concerns. We have already written elsewhere upon this topic (Tolmie et al, 2017) but there are notable axes along which journalistic work may vary: the time pressure in play and the presence of deadlines and temporal cycles; different organizational imperatives (such as what it is or isn't acceptable to publish); the presence of national and legal frameworks that stipulate what news coverage may look like (which can vary from country to country); editorial control (which can be highly idiosyncratic); the organisation of the local environment (for instance who sits with whom and the technical and physical resources available); the intended audience (which can vary significantly from news organisation to news organisation); and the kinds of story formats being developed (breaking news is very different to an in-depth article). The trick is to come up with a system that can manage this variability and still deliver what is required at the moment of use.

The nature of this problem was neatly summed up by one of the journalists in the following way:

> "It depends on context. Am I searching because I saw a newsflash? Perhaps I'm trying to get some perspectives. If I'm searching for that I will want eyewitness tweets and more in-depth reporting from trusted sources and not necessarily a reaction to what happened. But if I'm going in later on and we have already dealt with what happened, then probably I am looking for reactions. But when it's all a bit bundled it's hard to separate one from the other"

6 Conclusion

Both the PHEME Journalist Dashboard and Hercule are technologies that are already capable of being used to support journalistic work, even if still in rather restricted ways. This represents a significant amount of work and a significant amount of progress since the early prototypes were first evaluated in January 2016. It also represents a significant amount of cross-project collaboration, with most partners being implicated in the process in some way. In this report we have outlined the process whereby the dashboards have been evaluated, we have presented the results of those various evaluation exercises, and have examined the main implications of those results for the future development of the two technologies.

The main challenges that remain can be summarised in the following way:

- The provision of more overarching analytic overviews (preferably in graphic form). This is especially pressing for the PHEME Journalist Dashboard.
- Providing multilingual resources. This can be seen to apply to both dashboards.
- Improving the time taken to return results. Again this is especially important for the PHEME dashboard.
- Improving the interoperability between the dashboards and various social media platforms. This is particularly important for the PHEME dashboard.
- Improving the filtering and sorting functionality. Both dashboards have issues here.
- Improving the consistency and intelligibility of the indicators of veracity and controversiality. This is critical for the PHEME dashboard.
- Improving the ways in which results are titled and clustered. This is a major issue for both dashboards.
- Providing for flexible circumstances of use rather than adhering to a single primary use case. This again applies to both dashboards.

A further challenge that is not captured in the above list is an organisational one. Most of the journalists said they would use both of the dashboards right now if they were available. This reflects an ongoing strategy of making use of any technology at all that might offer an edge when it comes to processing potentially news-relevant information as quickly as possible and accomplishing potentially labour-intensive activities such as verification with as little overhead as possible. However, it was not clear that the journalists saw either dashboard as something that might offer them a more foundational solution to their information management needs. This poses questions regarding just what the scope of dashboards such as these should be. At present they often try to accommodate a wide range of functionality and, when it comes to providing something that is open to flexible circumstances of use, this would seem to be the correct way to proceed. However, actual use of the dashboards in an organisational context might situate them in a very different way, with them servicing just certain key aspects of the information assembly and verification process. This then remains as a topic for further investigation should one of the dashboards be brought into actual everyday working use.

7 Literature

1. Calcutt, A.& Hammond, P. (2011) *Journalism Studies: A Critical Introduction*. London: Routledge.
2. Crabtree, A., Rouncefield, M., & Tolmie, P. (2012). *Doing Design Ethnography*. London:

Springer Verlag.
3. Derczynski, L., Chester, S. & Bøgh, K.S. (2015) Tune Your Brown Clustering, Please. Proceedings of the conference on Recent Advances in Natural Language Processing (RANLP)
4. Diakopoulos, N., De Choudbury, M., & Naaman, M. (2012) Finding and Assessing Social Media Information Sources in the Context of Journalism. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '12), New York: ACM.
5. Harkin, J., Anderson, K., Morgan, L., and Smith, B. (2012) *Deciphering User-Generated Content in Transitional Societies*. Internews Report.
6. Koenemann-Belliveau, J., Carroll, J.M., Rosson, M.B., and Singley, M.K. (1994). Comparative usability evaluation: critical incidents and critical threads. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '94), New York: ACM, 245-251.
7. PHEME Project Deliverable D8.1 (2015) *Requirements Gathering, Use Case Design and Interface Mock-Ups*
8. PHEME Project Deliverable D8.3 (2016) *Digital Journalism Prototype*
9. PHEME Project Deliverable D8.3.1 (2017) *Digital Journalism Prototype (updated)*
10. Scriven, M. (1967). "The methodology of evaluation". In Stake, R. E. *Curriculum evaluation*. Chicago: Rand McNally. American Educational Research Association (monograph series on evaluation, no. 1
11. Singer, J.B. (2015) Out of Bounds: Professional Norms as Boundary Markers. In: M. Carlson & S.C. Lewis (Eds.), *Boundaries of Journalism: Professionalism, Practices and Participation*. Oxford: Routledge, 21-36.
12. Tolmie, P., Procter, R., Rouncefield, M., Liakata, M., and Zubiaga, A. (2015) Microblog Analysis as a Programme of Work. *arXiv preprint arXiv:1511.03193* (2015).
13. Tolmie, P., Procter, R., Randall, D., Rouncefield, M., Wong Sak Hoi, G., Burger, C., Liakata, M., and Zubiaga, A., (2017) Supporting the use of user generated content in journalistic practice. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '17), New York: ACM
14. Twidale, M., Randall, D., & Bentley, R. (1994). Situated evaluation of cooperative systems. *Proceedings of the Conference on Computer Supported Cooperative Work* (pp. 441-452). Chapel Hill: ACM.
15. Zubiaga, A., Liakata, A., Procter, R., Wong Sak Hoi, G. & Tolmie, P. (2016) Analysing How People Orient to and Spread Rumours in Social Media by Looking at Conversational Threads. *PLOS One, 11(3).*

8 Appendices

8.1 Appendix I – Instructions Provided to Participants for the First Formative Evaluation

First Formative Evaluation – 7 Jan 2016

Instructions:

- The PHEME tool is designed to give you access to developing stories on social media that are already, or may be set to become, rumours
- It is also designed to give you access to resources whereby you may be able to assess the validity or otherwise of the stories
- The aim of this session is to get you to run through an outline of your workflow using the PHEME tool and to provide us with feedback as you do so
- At this stage the tool is using canned rather than live data, but will be switching to using live social media feeds within the next few months
- You should approach the session as though you were currently working on the newsdesk (though it is possible for the tool to be used for feature writing as well)
- Please vocalize throughout anything you find problematic, are unsure about, or that strikes you as particularly useful
- You should also relate the exercise to other aspects of the workflow you would be engaged in at the same time: if you imagine you would probably work between the tool and other applications you can open up those other applications alongside and experiment with moving between them in the ways you'd expect to need to
- Please also bear in mind what you would do with other people within the various parts of the workflow and whether you feel the tool adequately supports interactions of that kind

Structure:

The session will be structured broadly along the lines of the previously observed workflow on the newsdesk, i.e.:

Scenario 1: Looking for stories

Scenario 2: Proposing stories as viable candidates

Scenario 3: Choosing specific stories to work on

Scenario 4: Verifying the information

Scenario 5: Actually writing the story

Scenario 6: Subbing the story (both passing it to be subbed and actually doing the subbing)

Scenario 7: Publishing the story using the publishing tool

8.2 Appendix II – Instructions Provided to Participants for the Second and Third Formative Evaluations

PHEME Second Formative Evaluation – 6 April 2016
PHEME Third Formative Evaluation – 7 September 2016

Instructions:

- As before, the PHEME tool is designed to give you access to developing stories on social media that are already, or may be set to become, rumours.
- It is also designed to give you access to resources whereby you may be able to assess the validity or otherwise of the stories.
- The aim of this session is to get you to try and use the PHEME tool to actually construct a story.
- It is now possible to access live data through the tool.
- However, you may be given specific other instructions on the day relating to workarounds etc. for known bugs or issues.
- As before, you should approach the session primarily as though you were currently working on the newsdesk (though it is possible for the tool to be used for feature writing as well).
- You should use whatever applications you would usually use for constructing a story alongside of the tool itself, as far as possible in the ways you would usually use those applications.
- You may also use a notepad to make notes if that is what you would usually do.
- Please vocalize throughout anything you find problematic, are unsure about, or that strikes you as particularly useful.

Structure:

You should try to follow the usual newsdesk workflow as far as you possibly can.

- First of all use the tool in the context of looking for viable stories.
- It is okay to identify several potential stories in the first instance.
- Then you should use the tool to further research stories you think you may want to develop more fully.

- After this we would like you to actually use the tool as part of preparing a short news story, through to the point where you would begin to think of publishing it.
- This can include the hunting out of any associated media (photos, related stories, infographics, maps, etc) that you think may be relevant.
- You may develop several stories alongside of one another if that seems sensible (e.g. if you spot a potential story but want to work on another one until more information becomes available).

- You will be given up to 2 hours to work on the task.
- The entire session will be recorded on video for subsequent close analysis, but the video material itself will not be used for any other purpose without your express permission.

8.3 Appendix III – Instructions & Questionnaire Provided to Participants for the Final Summative Evaluation

PHEME Journalist Dashboard - Summative Evaluation

General information and instructions:

This evaluation of the PHEME Journalist Dashboard is intended to be the final evaluation within the project. It therefore encompasses a broad range of tasks. There will also be a short evaluation of the fact-checking dashboard 'Hercule'.

- The PHEME tool is designed to give you access to developing stories on social media that are already, or may be set to become, rumours.
- It is also designed to give you access to resources whereby you may be able to assess the validity or otherwise of the stories.
- Hercule provides you with an additional resource whereby you may quickly assess the accuracy of facts present in various claims.
- Over the course of the evaluation you will be using a mixture of live and pre-annotated data.
- You may be given specific other instructions in some places relating to workarounds etc. for known bugs or issues.
- The entire session will be recorded on video for subsequent close analysis, but the video material itself will not be used for any other purpose without your express permission.

The evaluation is broken up into 4 parts:

Part 1:

Here you will be asked to work through a range of exercises using the dashboard and to provide feedback on a survey-type questionnaire.

This part of the evaluation should not take more than 45 minutes.

Part 2:

Here you will be asked to evaluate certain aspects of the fact-checking dashboard ‘Hercule’. Again you will be asked to provide responses on a questionnaire.

This part of the evaluation should not take more than 30 minutes.

Part 3:

Here you will be asked to use the dashboards you have tested to start to construct an actual news item.

This part of the evaluation should not take more than 45 minutes.

Part 4:

This part of the evaluation can be completed over the next few days. It asks you to provide us with some general observations about the PHEME journalist dashboard you have tested and your experience of using it.

This part of the evaluation will take about 30 minutes.

Before Part 1 begins you will be asked to undertake some preliminary activities in preparation for Parts 2 and 3.

Part 0

This part of the evaluation is designed to ensure that you will have some suitable materials to work with when you come to the other parts of the evaluation. As this will involve you interacting with the dashboards already we would also like some feedback on your experience of these tasks.

Task 0.0

Open the PHEME Journalist Dashboard page.

Use the Explore Social Media option to explore what kinds of new items are currently popular on Twitter and Reddit. Once you have done this search for two or three topics you would like to potentially research as news stories for later in the evaluation. Make sure that you choose fairly major ongoing international news events, keep your keywords simple, and enter them in a single box.

Task 0.1

Add each topic separately to create a new topic channel.

		not at all	not very	moderately	quite	very
A	How easy was it to understand the Explore Social Media feature?	**0**	**1**	**2**	**3**	**4**
B	How easy was it to search for topics?	**0**	**1**	**2**	**3**	**4**
C	How easy was it to add a topic?	**0**	**1**	**2**	**3**	**4**
Are there any comments you would like to add?						

Task 0.2

Now open the Hercule dashboard.

Examine the various features on the landing page to make sure they make sense to you.

Task 0.3

Add one or more topics to the dashboard (you may use the same topics as you did for the PHEME dashboard if you like).

		not at all	not very	moderately	quite	very
A	How easy was it to understand the various features?	**0**	**1**	**2**	**3**	**4**
B	Please make a note of any features you had to have explained to you.					
C	How easy was it to add a topic?	**0**	**1**	**2**	**3**	**4**
Are there any comments you would like to add?						

Your topic searches will now start to collect posts about the topics you are interested in whilst you are engaged in other tasks.

Part 1

The purpose of this part of the evaluation is to provide you with a series of structured tasks to encourage you to explore different features of the PHEME journalist dashboard. To enable us to control the dashboard responses you will be using pre-existing, pre-annotated data. Please enter feedback in the spaces provided. Please also try to provide commentary on your activity out loud as you proceed through the tasks. This part of the evaluation should not take longer than 45 minutes.

Task 1.1.1

Go to the PHEME journalist dashboard and explore Phemes in the topic Charlie Hebdo

Task 1.1.2

How many Phemes are there in this topic?

Answer:

		not at all	not very	moderately	quite	very
A	How easy was it to understand the notion of exploring phemes in a topic?	**0**	**1**	**2**	**3**	**4**
B	How easy was it to actually explore the phemes?	**0**	**1**	**2**	**3**	**4**
C	How easy was it to uncover how many phemes were in the topic?	**0**	**1**	**2**	**3**	**4**
D	How useful would the pheme exploration feature be to you in your work?	**0**	**1**	**2**	**3**	**4**
Are there any comments you would like to add?						

Task 1.2.1

Explore ways of filtering the Pheme list view

Task 1.2.2

Now explore ways of sorting the Pheme list view

		not at all	not very	moderately	quite	very
A	How easy was it to understand the filtering options?	**0**	**1**	**2**	**3**	**4**
B	Was the filtering easy to do?	**0**	**1**	**2**	**3**	**4**
C	Which filtering options would be most and least useful to you in your work?					
D	How easy was it to understand the sorting options?	**0**	**1**	**2**	**3**	**4**
E	Was the sorting easy to do?	**0**	**1**	**2**	**3**	**4**
F	Which ways of sorting would be most and least useful to you in your work?					
Are there any comments you would like to add?						

	not at all	not very	moderately	quite	very

Task 1.3.1

Which Pheme is the most active?

Answer:

Task 1.3.2

What is the most controversial Pheme?

Answer:

Task 1.3.3

Explore the most controversial Pheme. What does the controversiality score tell you about it?

Answer:

Task 1.3.4

What is the Pheme about?

Answer:

Task 1.3.5

How many images are contained in this Pheme?

Answer:

Task 1.3.6

How many verified authors are contained in this Pheme?

Answer:

		not at all	not very	moderately	quite	very
A	How easy is it to understand what the Pheme 'Activity' means?	**0**	**1**	**2**	**3**	**4**
B	How easy was it to locate which Pheme was most active?	**0**	**1**	**2**	**3**	**4**
C	How useful would being able to see Pheme 'Activity' be to you in your work?	**0**	**1**	**2**	**3**	**4**
D	How easy is it to understand what Pheme controversiality is about?	**0**	**1**	**2**	**3**	**4**
E	How easy was it to locate which Pheme was the most controversial?	**0**	**1**	**2**	**3**	**4**
F	How easy was it to understand the Pheme controversiality score?	**0**	**1**	**2**	**3**	**4**
G	How useful would Pheme controversiality be to you in your work?	**0**	**1**	**2**	**3**	**4**
H	How easy was it to understand what the Pheme was about?	**0**	**1**	**2**	**3**	**4**
I	How easy was it to locate how many images were in the Pheme?	**0**	**1**	**2**	**3**	**4**
J	How useful would knowing the number of images be to you in your work?	**0**	**1**	**2**	**3**	**4**
K	How easy was it to understand who counts as a verified author?	**0**	**1**	**2**	**3**	**4**
L	How easy was it to locate the number of verified authors in the Pheme?	**0**	**1**	**2**	**3**	**4**
M	How useful would knowing the number of verifiied authors be to you in your work?	**0**	**1**	**2**	**3**	**4**
Are there any comments you would like to add?						

Task 1.4.1

Choose a keyword and try using it to search the Phemes list, then save the results

		not at all	not very	moderately	quite	very
A	How easy was it to understand the keyword search feature?	**0**	**1**	**2**	**3**	**4**
B	How easy was it to use this feature?	**0**	**1**	**2**	**3**	**4**
C	How easy was it to save the results of your search?	**0**	**1**	**2**	**3**	**4**
D	How useful would the keyword search feature be to you in your work?	**0**	**1**	**2**	**3**	**4**
Are there any comments you would like to add?						

Task 1.5.1

Go to the detailed view of one of the controversial phemes.

Task 1.5.2

Does everything in the discussion belong in this Pheme?

Answer:

Task 1.5.3

What is the veracity label for this Pheme?

Answer:

Task 1.5.4

Does the veracity label make sense to you?

Answer:

		not at all	not very	moderately	quite	very
A	How easy was it to get to the detailed view of the Pheme?	**0**	**1**	**2**	**3**	**4**
B	How useful would this feature be to you in your work?	**0**	**1**	**2**	**3**	**4**
C	How easy was it to see why these threads were grouped together?	**0**	**1**	**2**	**3**	**4**
D	How easy is it to understand what a veracity label is telling you?	**0**	**1**	**2**	**3**	**4**
E	How easy was it to locate the veracity label?	**0**	**1**	**2**	**3**	**4**
F	How useful would a veracity label be to you in your work?	**0**	**1**	**2**	**3**	**4**
Are there any comments you would like to add?						

Task 1.6.1

Open one of the phemes

Task 1.6.2

Which URL(s) is linked to this Pheme?

Answer:

Task 1.6.3

Which image(s) is linked to it?

Answer:

Task 1.6.4

How close geographically are the authors to the event being discussed in this Pheme?

Answer:

		not at all	not very	moderately	quite	very
A	How easy is it to understand the notion of Linked URLs?	**0**	**1**	**2**	**3**	**4**
B	How easy was it to locate the Linked URLs?	**0**	**1**	**2**	**3**	**4**
C	How useful would this feature be to you in your work?	**0**	**1**	**2**	**3**	**4**
D	How easy was it to locate the images?	**0**	**1**	**2**	**3**	**4**
E	How useful would this feature be to you in your work?	**0**	**1**	**2**	**3**	**4**
F	How easy was it to find the location of the authors?	**0**	**1**	**2**	**3**	**4**
G	How useful would this feature be to you in your work?	**0**	**1**	**2**	**3**	**4**
Are there any comments you would like to add?						

Task 1.7.1

Try opening one of the conversations for the Pheme you were looking at above in Twitter.

Task 1.7.2

Now try opening one of the discussions in Reddit.

		not at all	not very	moderately	quite	very
A	How easy was it to open the conversation in Twitter?	**0**	**1**	**2**	**3**	**4**
B	How easy was it to open the discussion in Reddit?	**0**	**1**	**2**	**3**	**4**
C	How useful would these features be to you in your work?	**0**	**1**	**2**	**3**	**4**
Are there any comments you would like to add?						

Part 2

In this session you will be asked to explore and test out various features of the fact-checking dashboard 'Hercule', which currently assesses the accuracy of facts within claims posted on Twitter. As with Part 1 please enter feedback in the spaces provided and try to provide commentary on your activity out loud as you proceed through the tasks. This part of the evaluation should not take longer than 30 minutes.

2.1 Open the Landing Page for the Hercules dashboard once again and go to Topic Management.		
Do all of the features here make sense to you?.	**Yes**	**No**
If not, which ones do not?		
2.2 Click on 'Edit' for one of the topics and work through the available Topic Edit tabs.		
Is it clear to you what you can do here?.	**Yes**	**No**

If not, which features are not clear?		
2.3 Now got to the dashboard Results Panel.		
Do all of the features of this page make sense to you?.	**Yes**	**No**
If not, which features are not clear?		
2.4 Click on View Topic for one of the topics.		
Do all of the features here make sense?.	**Yes**	**No**
If not, which ones do not?		

2.5 What do you think ‘Checkworthy’ means?		
Does the idea of checkworthiness make sense to you?	**Yes**	**No**
Do you think this feature will be useful to you in your work?	**Yes**	**No**
Please explain the reasons for your answer.		
2.6 Now view one of the displayed stories in detail.		
Do all of the features make sense?.	**Yes**	**No**
If not, which ones do not?		
2.7 Now click on one of the concepts present in the story and explore the page this takes you to.		
Do all of the features make sense?.	**Yes**	**No**

If not, which ones do not?		
Do you think this feature will be useful to you in your work?	**Yes**	**No**
Please explain the reasons for your answer:		
Are there any resources missing here that you would like to see added in?		
2.8 Now go back to the Story View and click on 'Related World News'. Open one of the displayed articles.		
Was this feature easy to use?	**Yes**	**No**
Do you think it will be useful to you in your work?	**Yes**	**No**

Please explain the reasons for your answer.		
2.9 Go back once again to the Story View and open one of the tweets in Twitter.		
Was this easy to do?	**Yes**	**No**
Do you think it will be useful to you in your work?	**Yes**	**No**

2.10 How easy do you think the Hercule dashboard is to use overall (where 0 is not at all easy and 4 is extremely easy)	**0**	**1**	**2**	**3**	**4**
2.11 How well do you think it is laid out?	**0**	**1**	**2**	**3**	**4**
2.12 Are there any features you think could be presented more effectively?					
2.13 Are there any features you think are currently missing?					

2.14	Hopeless	Frustrating	Okay	Very Good	Excellent
How would you describe your overall experience of using the Hercule dashboard?					

2.15	Not at all	Not Very Well	Okay	Very Well	Perfectly
How well would using the dashboard fit into your usual workflow?					

2.16	Nowhere near as good	Not as good	About the same	Better	Much better
How does the dashboard compare with other tools you currently use?					

2.17 Would you use the dashboard in your everyday work if it were available?	**Yes**	**No**
Please explain your response in more detail:		

Part 3

- The aim of this session is to get you to try and use the PHEME journalist dashboard and Hercule as you might in your actual working practice. To this end you will be using live data.
- You should approach the session primarily as though you were currently working on the newsdesk (though it is possible for the dashboards to be used for feature writing as well).
- You should use whatever applications you would usually use for constructing a story alongside of the dashboards themselves, as far as possible in the ways you would usually use those applications.
- You may also use a notepad to make notes if that is what you would usually do.
- Please vocalize throughout anything you find problematic, are unsure about, or that strikes you as particularly useful.

Steps:

- First of all use the dashboards in the context of looking for viable stories. You can do this by drawing upon the topics you added to the dashboards at the beginning of the evaluation. It is okay to identify several potential stories in the first instance.
- Then you should use the dashboards to further research stories you think you may want to develop more fully.
- If possible we would like you to actually use the dashboards as part of preparing a short news story, through to the point where you would begin to think of actually writing it.
- This can include the hunting out of any associated media (photos, related stories, infographics, maps, etc) that you think may be relevant.
- You may develop several stories alongside of one another if that seems sensible (e.g. if you spot a potential story but want to work on another one until more information becomes available).

Part 4

This part of the evaluation relates exclusively to the **PHEME journalist dashboard.** Our aim here is to get you to give us your overall impressions of how it compares to other resources that you use. As with Task 1 it is constructed as a questionnaire, though we would encourage you to make use of the free text boxes as well wherever possible. You may complete this questionnaire any time up until the end of the day on Tuesday 7th March.

4.1	Under 24	25-34	35-44	45-54	55 or Over
How old are you?					

4.2	Under 5 years	Between 6 & 10 years	Between 11 & 15 years	Between 16 & 20 years	21 years or more
How long have you been a journalist?					

4.3 Have you always worked in news organisations?	**Yes**	**No**
If the answer is no, what other kinds of organisations have you worked in?		

4.4 Have you ever worked as a freelance journalist?	**Yes**	**No**

4.5	Never or very rarely	Occasionally	Every couple of weeks	Weekly	Daily

4.5	Never or very rarely	Occasionally	Every couple of weeks	Weekly	Daily
How often do you work on the newsdesk or work shifts to produce quick stories?					

4.6	Less than 10%	Less than 50%	About 50%	More than 50%	90% or more
What percentage of your overall work would you say is newsdesk/quick story work?					

4.7	Not at all	A bit	Moderately	Very	I'm an Expert
How expert are you in using social media?					

4.8	
What kinds of social media do you use in your work?	

4.9	Never or very rarely	Occasionally	A moderate amount	A lot	All the time
How often do you use social media in your work?					

4.10 Which social media platforms do you use the most?	

4.11 Please rate how easy the dashboard was to use overall (where 0 is not at all easy and 4 is extremely easy)	**0**	**1**	**2**	**3**	**4**

4.12 Please rate how well laid out you considered the dashboard to be	**0**	**1**	**2**	**3**	**4**

4.13 Please rate the overall look and feel of the dashboard	**0**	**1**	**2**	**3**	**4**

4.14 Are there any features you think could be presented more effectively?	

4.15 Are there any features you think are currently missing?	

4.16 Please rate how easy you found it to be to understand all of the different features	**0**	**1**	**2**	**3**	**4**

4.17 What features of the dashboard would be most useful to you in your work as a journalist?	

4.18 What features of the dashboard would be the *least* useful to you in your work?	

4.19 What features of the dashboard did you have the most trouble with?	

4.20 Did you encounter any performance issues with the dashboard?	**Yes**	**No**
If so, what?		

4.21	Hopeless	Frustrating	Okay	Very Good	Excellent
How would you describe your overall experience of using the dashboard?					

4.22	Not at all	Not Very Well	Okay	Very Well	Perfectly
How well would using the dashboard fit into your usual workflow?					

4.23	Nowhere near as good	Not as good	About the same	Better	Much better
How does the dashboard compare with any other social media tools you are using?					

4.24	Nowhere near as good	Not as good	About the same	Better	Much better
How does the dashboard compare with other tools you use more generally (such as the newswires)?					

4.25 Would you use the dashboard in your everyday work if it were available?	**Yes**	**No**
Please explain your response in more detail.		

4.26						
For each of the following activities please rate the potential usefulness of the dashboard (where 0 is not at all useful and 4 is extremely useful)						
A	Newsdesk Work	**0**	**1**	**2**	**3**	**4**
B	Writing Feature articles	**0**	**1**	**2**	**3**	**4**
C	Freelancing	**0**	**1**	**2**	**3**	**4**