# An Infrastructure for Empowering Internet Users to Handle Fake News and Other Online Media Phenomena

Online media and digital communication technologies have an unprecedented, even increasing level of social, political and also economic relevance. This article proposes an infrastructure to address phenomena of modern online media production, circulation and manipulation by establishing a distributed architecture for automatic processing and human feedback.

This article addresses key challenges of the digital age (Sect. 2) by introducing and proposing the vision of a technological infrastructure (Sect. 3); the concept has been devised in a research and technology transfer project, in which smart technologies for curating large amounts of digital content are being developed and applied by companies that cover different sectors including journalism (Rehm and Sasaki 2015; Bourgonje et al. 2016a,b; Rehm et al. 2017). Among others, we currently develop services aimed at the detection and classification of abusive language (Bourgonje et al. 2017a) and clickbait content (Bourgonje et al. 2017b). The proposed hybrid infrastructure combines automatic language technology components and user-generated annotations and is meant to empower internet users better to handle the modern online media phenomena ...

## *Contents*

# Georg Rehm (GM)

## Description:

DFKI GmbH, Language Technology Lab, Alt-Moabit 91c, 10559 Berlin, Germany

### Stakeholder(s):

**Online Media** :
*The umbrella term "fake news" is often used to refer to a number of different phenomena around online media production, circulation, reception and manipulation that emerged in recent years and that have been receiving a lot of attention from multiple stakeholders including politicians, journalists, researchers, non-governmental organisations, industry and civil society. In addition to the challenge of dealing with "fake news", "alternative facts" as well as "post-truth politics", there is an increasing amount of hate speech, abusive language and cyber bullying taking place online.*

**Politicians** :
*Among the interested stakeholders are politicians who have begun to realise that, increasingly, major parts of public debates and social discourse are carried out online, on a small number of social networks.*

**Social Networks** :
*We have witnessed that not only online discussions but also the perception of trends, ideas, theories, political parties, individual politicians, elections and societal challenges can be subtly influenced and significantly rigged using targeted social media campaigns, devised at manipulating opinions to create long-term sustainable mindsets on the side of the recipients. We live in a time in which online media, online news and online communication have an unprecedented level of social, political and economic relevance.*

**Social Science Researchers** :
*Due to the intrinsic danger of successful large-scale manipulations the topic is of utmost importance. Many researchers from the Social Sciences and Computer Science currently work on the topic.*

**Computer Science Researchers** :
*An idea often mentioned is to design, develop and deploy technologies to improve the situation, maybe even to solve it altogether, thanks to recent breakthroughs in AI (Metz 2016; Gershgorn 2016; Martinez-Alvarez 2017; Chan 2017), while at the same time not putting in place a centralised infrastructure, which could be misused for censorship, manipulation or mass surveillance.*

**HTML5 Users** :
*Technically, online content is predominantly consumed through two possible channels, both of which rely substantially on World Wide Web technology and established web standards. Users either read and interact with content directly on the web (mobile or desktop versions of websites) or through dedicated mobile apps; this can be considered using the web implicitly as many apps make heavy use of HTML5 and other web technologies. The World Wide Web itself still is and, for the foreseeable future, will continue to be the main transport medium for online content.*

**Worldwide Web Users** :
*The infrastructure suggested by this article is, hence, designed as an additional layer on top of the World Wide Web. The scope and ambition of the challenge is immense because the infrastructure needs to be able to cope with millions of users, arbitrary content types, hundreds of languages and massive amounts of data. Its goal is to empower users by enabling them to balance out network and filter bubble effects and to provide mechanisms to filter for abusive content.*

**Analysis Services** :
*The burden of analysing and fact checking online content is often shifted to the reader (Sect. 2), which is why corresponding analysis and curation services need to be made available in an efficient and ubiquitous way.*

**Curation Services** :
*The services need to be designed to operate in and with the web stack of technologies, they need to support users in their task of reading and curating content within the browser in a smarter and, eventually, more balanced way. This can be accomplished by providing additional, also alternative opinions and view points, by presenting other, independent assessments, or by indicating if content is dangerous, abusive, factual or problematic in any way. Fully automatic technologies (Rubin et al. 2015; Schmidt and Wiegand 2017; Horne and Adal 2017; Martinez-Alvarez 2017) can take over a subset of these tasks but, given the current state of the art, not all, which is why the approach needs to be based both on simple and complex automatic filters and watchdogs as well as human intelligence and feedback.*

**Content Consumers** :
*The same tools to be used by content consumers can and should also be applied by content creators, e.g., journalists and bloggers. Those readers who are interested to know more about what they are currently reading should be able to get the additional information as easily as possible, the same applies to those journalists who are interested in fact-checking the content they are researching for the production of new content. Readers of online content are users of the World Wide Web. They need, first and foremost, web-based tools and services with which they can process any type of content to get additional information on a specific piece, be it one small comment on a page, the main content component of a page (for example, an article) or even a set of interconnected pages (one article spread over multiple pages), for which an assessment is sought.*

**Journalists**

**Bloggers**

**CMC Researchers** :
*Related Work ~ Research on Computer-Mediated Communication (CMC) has a long tradition.*

*— continued next page*

*Stakeholders (continued)*

**Scholars** :
*Scholars initially concentrated on different types of communication media such as e-mail, IRC, Usenet newsgroups, and different hypertext systems and document types, especially personal home pages, guestbooks and, later, discussion fora (Runkehl et al. 1998; Crystal 2001; Rehm 2002). Early on, researchers focused upon the (obvious) differences between these new forms of digital communication and traditional forms, especially when it comes to linguistic phenomena that can be observed on the text surface (smileys, emoticons, acronyms etc.). Several authors pointed out that the different forms of CMC have a certain oral and spoken style, quality and conceptualisation to them, as if produced spontaneously in a casual conversation, while being realised in a written medium (Haase et al. 1997).*

**Global Population** :
*If we now fast forward to 2017, a vastly different picture emerges. About half of the global population has access to the internet, most of whom also use the World Wide Web and big social networks. Nowadays the internet acts like an amplifier and enabler of social trends. It continues to penetrate and to disrupt our lives and social structures, especially our traditions of social and political debates.*

**Online Media** :
*The relevance of online media, online news and online communication could not be any more crucial.*

**Social Communities** :
*While early analyses of CMC, e.g., (Reid 1991), observed that the participants were involved in the "deconstruction of boundaries" and the "construction of social communities", today the exact opposite seems to be the case: not only online but also offline can we observe the trend of increased, intricately orchestrated, social and political manipulation, nationalism and the exclusion of foreigners, immigrants and seemingly arbitrary minorities – boundaries are constructed, social communities deconstructed, people are manipulated, individuals excluded.*

**Text Analyzers** :
*There is a vast body of research on the processing of online content including text analytics (sentiment analysis, opinion and argument mining), information access (summarisation, machine translation) and document filtering (spam classification), see (Dale 2017). Attempting to classify, among others, the different types of false news shown in Table 1 requires, as several researchers also emphasise, a multi-faceted approach that includes multiple different processing steps.*

**Sentiment Analyzers**

**Opinion Miners**

**Argument Miners**

**Document Filtering Services**

**Propaganda Detectives** :
*We have to be aware of the ambition, though, as some of the "fake news detection" use case scenarios are better described as "propaganda detection", "disinformation detection", maybe also "satire detection". These are difficult tasks at which even humans often fail. Current research in this area is still fragmented and concentrates on very specific sub-problems, see, for example, the Fake News Challenge, the Abusive Language Workshop, or the Clickbait Challenge.7 What is missing, however, is a practical umbrella that pulls the different pieces and resulting technology components together and that provides an approach that can be realistically implemented and deployed including automatic tools as well as human annotations.*

**Disinformation Detectives**

**Satire Detectives**

**German Federal Ministry of Education and Research** :
*Acknowledgments. The author would like to thank the reviewers for their insightful comments and suggestions. The project "Digitale Kuratierungstechnologien" (DKT) is supported by the German Federal Ministry of Education and Research (BMBF), "Unternehmen Region", instrument Wachstumskern-Potenzial (no. 03WKP45). More information: http://www.digitale-kuratierung.de.*

# Vision

Internet users are empowered to handle online media phenomena

# Mission

To address phenomena of modern online media production, circulation and manipulation

# Values

**Effectiveness**: In order for these tools and services to work effectively, efficiently and reliably, they need to have several key characteristics, which are critical for the success of the approach.

**Efficiency**

**Reliability**

**Federation**: Like the Internet and the World Wide Web, the proposed infrastructure must be operated in a federated, i. e., de-centralised setup – a centralised approach would be too vulnerable for attacks or misuse.

**Decentralisation**

**Standardisation**: Any organisation, company, research centre or NGO should be able to set up, operate and offer services (Sect. 3.1) and pieces of the infrastructure. The internal design of the algorithms and tools may differ but their output should comply to a standardised metadata format (MGM).

**Personalization**: It is rather likely that political biases in different processing models meant to serve the same purpose cannot be avoided, which is especially likely for models based on large amounts of data, which, in turn, may inherently include a political bias. This is why users must be enabled to configure their own personalised set of tools and services to get an aggregated value, for example, with regard to the level of hate speech in content or its political bias.

**Combinability**: Services and tools must be combinable, i.e., they need to comply to standardised input and output formats (Babakar and Moy 2016).

**Transparency**: They also need to be transparent (Martinez-Alvarez 2017).

**Documentation**: Only transparent, i.e., fully documented, checked, ideally also audited approaches can be trustworthy

**Auditability**

**Trustworthiness**

**Universality**: Access to the infrastructure should be universal and available everywhere, i.e., in any browser, which essentially means that, ideally, the infrastructure should be embedded into the technical architecture of the World Wide Web.

**Accessibility**: As a consequence, access mechanisms should be available in every browser, on every platform, as native elements of the GUI.

**Unobtrusiveness**: These functions should be designed in such a way that they support users without distracting them from the content.

**Scalability**: Only if the tools are available virtually anywhere, can the required scale be reached.

**Configurability**: The user should be able to configure and to combine multiple services, operated in a de-centralised way, for a clearly defined purpose in order to get an aggregated value. There is a danger that this approach could result in a replication and shift of the filter bubble effect (Sect. 2) onto a different level but users would at least be empowered actively to configure their own personal set of filters to escape from any resulting bubble. The same transparency criterion also applies to the algorithm that aggregates multiple values.

## Online Media & Communications

*Enable automatic processing and human feedback on online media and communications phenomenon.*

### Stakeholder(s)

**Curation Services** :
*The curation services should be thought of as an inherent technology component of the World Wide Web, for which intuitive and globally acknowledged user-interface conventions can be established, such as, for example, traffic light indicators for false news content (green: no issues found; yellow: medium issues found and referenced; red: very likely false news). Table 2 shows a first list of tools and services that could be embedded into such a system. Some of these can be conceptualised and implemented as automatic tools (Horne and Adal 2017), while others need a hybrid approach that involves crowd-sourced data and opinions. In addition to displaying the output of these services, the browser interface needs to be able to gather, from the user, comments, feedback, opinions and sentiments on the current piece of content, further to feed the crowd-sourced data set. The user-generated data includes both user-generated annotations (UGA) and also user-generated metadata (UGM). Automatically generated metadata are considered machine-generated metadata (MGM).*

**Curation Tools & Services** :
*A fully automatic solution would work only for a very limited set of cases. A purely human-based solution would work but required large amounts of experts and, hence, would not scale. This is why we favour, for now, a hybrid solution. | This list is meant to be indicative rather than complete. For example, services for getting background information on images are not included (Gupta et al. 2013). Such tools could help pointing out image manipulations or that an old image was used, out of context, to illustrate a new piece of news.*

**Political Bias Indicator Services** :
*Tool or service: Political bias indicator ~ Indicates the political bias (Martinez-Alvarez 2017) of a piece of content, e.g., from far left to far right. | Approach: Automatic*

**Hate Speech Indicator Services** :
*Tool or service: Hate speech indicator ~ Indicates the level of hate speech a certain piece of content contains. | Approach: Automatic*

**Reputation Indicator Services** :
*Tool or service: Reputation indicator ~ Indicates the reputation, credibility (Martinez-Alvarez 2017), trustworthiness, quality (Filloux 2017) of a certain news outlet or individual author of content. | Approach: Crowd, automatic*

**Fact Checking Services** :
*Tool or service: Fact checker ~ Checks if claims are backed up by references, evidence, established scientific results and links claims to the respective evidence (Babakar and Moy 2016) | Approach: Automatic*

**Fake News Indicator Services** :
*Tool or service: Fake news indicator ~ Indicates if a piece of content contains non-factual statements or dubious claims (Horne and Adal 2017; Martinez-Alvarez 2017) | Approach: Crowd, automatic*

**Opinion Inspection Services** :
*Tool or service: Opinion inspector ~ Inspect opinions and sentiments that other users have with regard to this content (or topic) – not just the users commenting on one specific site but all of them. | Approach: Crowd, automatic*

**Language Technology Researchers** :
*Building Blocks of the Proposed Infrastructure ~ Research in Language Technology and NLP currently concentrates on smaller components, especially watchdogs, filters and classifiers (see Sect. 4) that could be applied under the umbrella of a larger architecture to tackle current online media phenomena (Sect. 2). While this research is both important and crucial, even if fragmented and somewhat constrained by the respective training data sets (Rubin et al. 2015; Conroy et al. 2015; Schmidt and Wiegand 2017) and limited use cases, we also need to come to a shared understanding how such components can be deployed and made available. The proposed infrastructure consists of several building blocks ...*

The tools and services should be available to every web user without the need of installing any additional third-party software. This is why the services, ideally, should be integrated into the browser on the same level as bookmarks, the URL field or the navigation bar, i.e., without relying on the installation of a plugin.

### Building Block 1. Architecture

*Embed the approach into the architecture of the World Wide Web itself.*

Natively embedded into the World Wide Web – An approach that is able to address modern online media and communication phenomena adequately needs to operate on a web-scale level. It should natively support cross-lingual processing and be technically and conceptually embedded into the architecture of the World Wide Web itself... Only if all users have immediate access to the tools and services suggested in this proposal can they reach its full potential. The services must be unobtrusive and cooperative, possess intuitive usability, their

recommendations and warnings must be immediately understandable, it must be simple to provide general feedback (UGM) and assessments on specific pieces of content (UGA).

**Stakeholder(s):**

**Browser Vendors** :
*It should be standardised, endorsed and supported not only by all browser vendors but also by all content and media providers, especially the big social networks and content hubs.*

**Content Providers**

**Media Providers**

**Big Social Networks**

**Content Hubs**

## Building Block 2. Annotations

*Apply the Web Annotations standard.*

Web Annotations – Several pieces of the proposed infrastructure are already in place... Web Annotations are the natural mechanism to enable users and readers interactively to work with content, to include feedback and assessments, to ask the author or their peers for references or to provide criticism. The natural language content of Web Annotations (UGA) can be automatically mined using methods such as sentiment analysis or opinion mining – in order to accomplish this across multiple languages, cross-lingual methods need to be applied (Rehm et al. 2016). However, there are still limitations... Federated sets of annotation stores or repositories are not yet foreseen, neither are native controls in the browser that provide aggregated feedback, based on automatic (MGM) or manual content assessments (UGM, UGA).

**Stakeholder(s):**

**World Wide Web Consortium** :
*One key component are Web Annotations, standardised by the World Wide Web Consortium (W3C) in early 2017 (Sanderson et al. 2017a,b; Sanderson 2017). They enable users to annotate arbitrary pieces of web content, essentially creating an additional and independent layer on top of the regular web. Already now Web Annotations are used for multiple individual projects in research, education, scholarly publishing, administration and investigative journalism.4*

**Content Providers** :
*Content providers need to enable Web Annotations by referencing a corresponding JavaScript library.*

**Social Networks** :
*Another barrier for the widespread use and adoption of Web Annotations are proprietary commenting systems, as used by all major social networks.*

**Hypothes.is** :
*Nevertheless, services such as Hypothes.is enable Web Annotations on any web page, but native browser support, ideally across all platforms, is still lacking. A corresponding browser feature needs to enable both free-text annotations of arbitrary content pieces (UGA) but also very simple flagging of problematic content, for example, "content pretends to be factual but is of dubious quality" (UGM). Multiple UGA, UGM or MGM annotations could be aggregated and presented to new readers of the content to provide guidance and indicate issues.*

## **Building Block 3.** Metadata

*Agree upon a standard metadata schema*

Metadata Standards – Another needed piece of the architecture is an agreed upon metadata schema, i.e., a controlled vocabulary, (Babakar and Moy 2016) to be used both in manual annotation scenarios (UGM) and also by automatic tools (MGM). Its complexity should be as low as possible so that key characteristics of a piece of content can be adequately captured and described by humans or machines. With regard to this requirement, W3C published several standards to represent the provenance of digital objects (Groth and Moreau 2013; Belhajjame et al. 2013a). These can be thought of as descriptions of the entities or activities involved in producing or delivering a piece of content to understand how data was collected, to determine ownership and rights or to make judgements about information to determine whether to trust content (Belhajjame et al. 2013b).

**Stakeholder(s):**

**Schema.org** :
*An alternative approach is for publishers to use Schema.org's ClaimReview5 markup after specific facts have been checked.*

**W3C** :
*The needed metadata schema can be based on the W3C provenance ontology and/or Schema.org. Additional metadata fields are likely to be needed.*

## **Building Block 4.** Tools & Services

*Enable readers to use Web Annotations to provide comments and results of factual research on online content.*

Tools and Services – Web Annotations can be used by readers of online content to provide comments or to include the results of researched facts (UGA, UGM). Automatic tools and services that act as filters or watchdogs can make use of the same mechanisms (MGM, see Sect. 3.1).

**Stakeholder(s):**

**Watchdogs** :
*These could be functionally limited classifiers, for example, regarding abusive language, or sophisticated NLU components that attempt to check certain statements against one or more knowledge graphs. Regardless of the complexity and approach, the results can be made available as globally accessible Web Annotations (that can even, in turn, be annotated themselves). Services and tools need to operate in a decentralised way, i.e., users must be able to choose from a wide variety of automatic helpers. These could, for example, support users to position content on the political spectrum, either based on crowd-sourced annotations, automatic tools, or both (see Table 2).*

## **Building Block 5.** Decentralisation

*Federate and decentralise the infrastructure.*

Decentralised Repositories and Tools – The setup of the infrastructure must be federated and decentralised to prevent abuse by political or industrial forces. Data, especially annotations, must be stored in decentral repositories, from which browsers retrieve, through secure connections, data to be aggregated and displayed (UGM, UGA, MGM, i.e., annotations, opinions, automatic processing results etc.). In the medium to long term, in addition to annotations, repositories will also include more complex data, information and knowledge that tools and services will make use of, for example, for fact checking... Already now we can foresee more sophisticated methods of validating and fact-checking arbitrary pieces of content using systems that make heavy use of knowledge graphs, for example, through automatic entity recognition and linking, relation extraction, event extraction and mapping etc. One of the key knowledge bases missing, in that regard, is a Web Annotation-friendly event-centric knowledge graph, against which fact-checking algorithms can operate.6 Basing algorithms that are supposed to determine the truth of an arbitrary statement on automatically extracted and formally represented knowledge creates both practical and philosophical questions, among others, who checks these automatically extracted knowledge structures for correctness? How do we represent conflicting view points and how do algorithms handle conflicting view points when determining the validity of a statement? How do we keep the balance between multiple subjective opinions and an objective and scientific ground-truth?

**Stakeholder(s):**

**Knowledge Graphs** :
*In parallel to the initiative introduced in this article, crowd-sourced knowledge graphs such as Wikidata or DBpedia will continue to grow ...*

**Wikidata**

**DBpedia**

**Semantic Databases** :
*... the same is true for semantic databases such as BabelNet and many other data sets, usually available and linkable as Linked Open Data.*

**BabelNet**

## **Building Block 6.** Aggregation

*Aggregate annotations.*

Aggregation of Annotations – The final key building block of the proposed infrastructure relates to the aggregation of automatic and manual annotations, created in a de-centralised and highly distributed way by human users and automatic services (UGA, UGM, MGM). Already now we can foresee very large numbers of annotations so that the aggregation and consolidation will be a non-trivial technical challenge. This is also true for those human annotations that are not based on shared metadata vocabularies but that are free text – for these free and flexible annotations, robust and also multilingual annotation mining methods need to be developed.

## Administrative Information

**Start Date:**
**End Date:**

**Publication Date:** **2021-09-05**
**Source:** https://link.springer.com/content/pdf/10.1007%2F978-3-319-73706-5_19.pdf

### Submitter:

**Given Name:** Owen
**Surname:** Ambur
**Email:** Owen.Ambur@verizon.net
**Phone:**