

# Semantic Web Applications for Danish News Media

Astrid Ildor

Department of Design and Communication, University of Southern Denmark,  
Universitetsparken 1, 6000 Kolding, Denmark

**Keywords:** Semantic Web, Linked Data, Web 3.0, News Media, Journalism, RDF, Semantic Annotation, URI, Web Ontologies, Application Development.

**Abstract:** Most news media possess a publish-and-forget mindset: Once a news article is published, the information it contains devalue in the messy place of the unstructured Web and great potentials of combining and reusing data is missed. News media has long constituted an area of interest for Semantic Web researchers, but few studies merge technical knowledge with editorial insights. To fill the gap in literature, this study combines technical analysis with interviews and Participatory Design studies with eight Danish news journalists and digital editors. The exploration reveals three areas within the journalistic work process with significant potential of improvement: Journalists' challenge of finding the right person to comment on a specific topic, issues of finding previously published articles, and the need for generating infoboxes. Each area is examined as a type of Semantic Web application. It is demonstrated how profound annotation of persons, places, organisations, and key terms mentioned in a body of articles is required for each application. Trustworthiness is another major challenge as this cannot yet be fully achieved within the concept of Semantic Web.

## 1 INTRODUCTION

News media publish trustworthy information on an hourly basis; however, the vast majority of information is not being archived or annotated for the purpose of reuse. Great potentials of combining different datasets or autogenerating new information based on accessible data cannot be fulfilled as lots of data – including the content of most news articles – is not machine-readable. To attain machine-readability, Semantic Web (SW) reuses the Web's global indexing and naming scheme, meaning that every semantic concept can be annotated a unique identifier (Domingue et al., 2011) – such data is known as *linked data*. Through formal computational ontologies and ontology-aware technologies, relationships between concepts published on the SW can be processed and understood within specific domains. In this context, SW contains potential as a distributed reasoning machine that not only can execute extremely precise searches but also analyse existing data and create new knowledge (Goddard, & Byrne, 2010). This can potentially improve the process of any news journalist and entail innovative storytelling and knowledge mediation.

However, before media organisations can be expected to make large investments in archives of linked data, journalists and other stakeholders must be convinced that there are costly problems associated with their current suite of processes and technologies, and that SW applications can help solve these. On the other hand, news media has long constituted an area of interest for SW researchers, but remarkable little research include insights of editors and journalists. To fill the gap in literature, this study attempts to answer the following research question: *How can the work process of news journalists and the user experience of news journalism be improved in the context of SW?*

The problem background is discussed in Section 2 on the basis of a small literature review carried out as part of the study. The importance of linked data and Resource Description Framework (RDF) is then commented in Section 3, before Section 4 sets out the method of the qualitative study. Qualitative analysis reveals three areas within Danish news journalism with significant potential of improvement in the context of SW. These areas are discussed as three types of SW applications for the news media industry in Section 4.1–4.3. Finally, Section 5 sums up the study in a short discussion and conclusion.

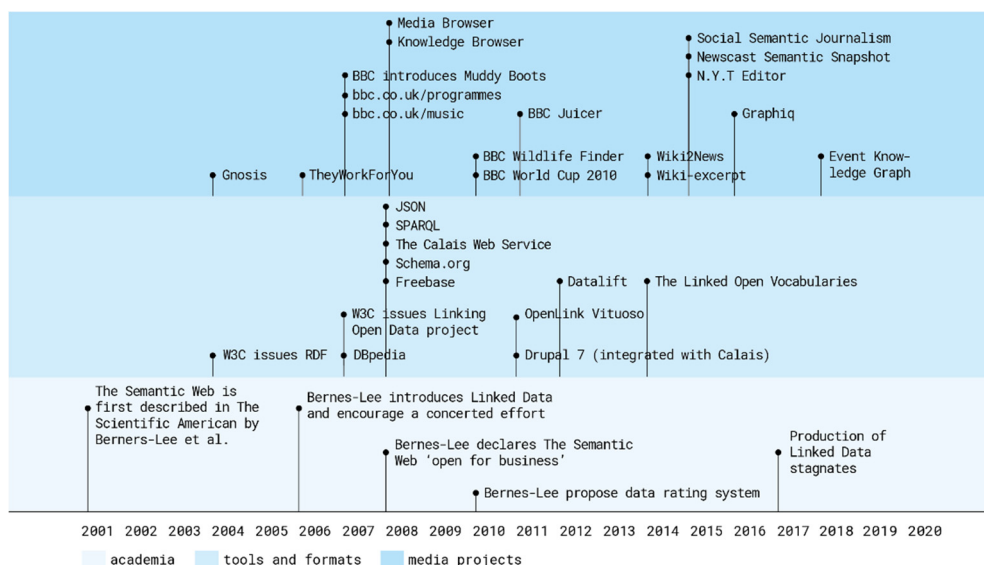


Figure 1: Semantic Web applications for news media.

## 2 PROBLEM BACKGROUND

In February 2008, semantic annotation tools reached a point of development where Berners-Lee and the W3C declared SW open for business (Miller, 2008). Coinciding, multiple news media organisations launched initiatives for annotating their archives of news articles in order to generate SW applications (see Figure 1): In 2008, Thomson Reuters launched a Web service capable of extracting entities and relationships in text documents and annotating these with linked data URIs (Hendler et al., 2011). In cooperation with Rattle Research, the BBC has developed a similar service to empower applications such as BBC Wildlife Finder which repurposes data from Wikipedia, WWF, and the IUCN’s Red List of Threatened Species and combines it with natural world footage from the BBC archive (Raimond, Scott, Oliver, Sinclair, & Smethurst, n.d.). Finally, in 2015, New York Times (N.Y.T.) launched a semi-automated annotation tool trained to apply semantic N.Y.T.-resources to plain text (N.Y.T. Labs, 2015). N.Y.T. has annotated their archive of articles from 1981 to today and on this basis generated multiple SW applications. A handful other smaller news media

organisations and researchers have launched or described similar SW applications – in total 24 SW applications or descriptions of thought-up applications for the news media industry have been identified as part of a literature review for this study. Analysis of these demonstrates that none of the applications seem to be based on systematic empirical insights of news journalists or publishers.

A majority of the applications are concerned with content search and new ways of presenting already published information. Only five<sup>1</sup> of the 24 applications are identified as *research tools to support journalists’ work processes*. In comparison, all of the applications proposed in this study can be characterised as research tools which in some way support the work process of Danish news journalists. This comparison indicates, that it might be beneficial to change the objective of SW application development for the news media industry: Instead of focusing on how already published information can be presented, this study demonstrates that potentials of SW are more likely to be unfolded within news journalists’ work processes.

<sup>1</sup> The five examples identified as research tools for journalists are: **Wiki2News** and **Wiki-excerpt** (Rudnik et al., 2019). **Social Semantic Journalism** (Heravi & McGinnis, n.d.). The last two examples are thought-up scenarios where Semantically-Interlinking Online Communities (**SIOC**) are applied to combine data from

different sources about a politician relevant for a specific news article (Raimond et al., n.d.). In the other example SIOC is applied to research the term ‘bogsnoorkelling’ across Facebook, Technorati, Flickr, and YouTube (Meek, 2008).

### 3 THE CONCEPT OF LINKED DATA

In order for the Web of linked documents to evolve into a Web of linked data, Berners-Lee introduced the Linked Data Principles as a best practice for publishing structured data on the Web:

1. Use URIs as names for things
2. Use HTTP URIs so that people can look up those names
3. When someone looks up an URI, provide useful information, using the standards RDF and SPARQL
4. Include links to other URIs, so that they can discover more things (Berners-Lee, 2006)

The first principle advocates using URIs to identify, not just Web documents, but also real-world objects and abstract concepts. Unicode in combination with URI extends support for identifying any type of resource regardless of its text and scripting language (Alam et al., 2015).

The second principle advocates the use of HTTP URIs to identify objects and abstract concepts just like HTTP is the universal access mechanism for the traditional Web (Berners-Lee, 2006).

The third principle advocates that HTTP clients should be able to look up any URI and retrieve a description of the resource. The agreement of HTML as the dominant document format has been crucial for the Web's ability to scale (Hendler et al., 2011). Similarly, it is important to agree on a standardised content format for URI descriptions. According to the third principle, this format should be Resource Description Framework (RDF) which is readable for both humans and machines (Berners-Lee, 2006). It is common practice to use different URIs to identify real-world object and the document that describes it.

The fourth principle advocates the use of links to connect not only Web documents, but any concept described on the Web (Berners-Lee, 2006). Such links in a linked data context are called RDF links.

#### 3.1 Resource Description Framework

RDF is a data model for publishing statements on the Web. Each statement is represented as a *triple* consisting of a subject, a predicate, and an object. The subject is the URI identifying the described resource. The object can either be a simple *literal value* or the URI of another resource that is somehow related to the subject. The predicate is also identified by an URI and describes the relationship between subject and object. RDF provides a data model for describing resources, but it does not provide any

domain-specific terms for describing classes of things, and how they might relate. This function is served by lightweight RDFS ontologies, thus predicate-URIs come from standard vocabularies or ontologies. The strength of RDF lies in the flexibility of integration. RDF graphs can quite easily be merged by sharing particular resources, or claiming two resources to be the same, although their identifier might be different (Domingue et al., 2011).

Resource Description Framework in Attributes (RDFa) is a serialisation format that embeds RDF triples in the HTML document, meaning that existing content within the HTML code can be annotated with RDFa (Hendler et al., 2011).

RDF and the Linked Data Principles allow for publishing and accessing simple facts but do not support more complex queries (Hendler, Heath, & Bizer, 2011). To retrieve this type of information, SPARQL Protocol and RDF Query Language (SPARQL) can be applied. The query language is designed for evaluating queries against RDF datasets and to ask meaning-driven questions to databases of structured data on the Web (Wood, Marsha, Luke, & Hausenblas, 2013).

#### 3.2 Research on Proof and Trust

In most SW applications, it is fundamental to know where presented information comes from and how resulting conclusions have been constructed. Technologies for including automatic proof checking are however not yet standardised, and research lacks to provide answers to several questions.

In SW applications, technologies of *unifying logic* operate on top of the ontology to make new inferences (Pandey & Sanjay, 2010). However, standards to ensure transparency on how applied ontologies and reasoning mechanisms are constructed are still missing and require more research. Without transparency, biased or manipulated ontologies can provide answers which cannot be distinguished as true or false.

Berners-Lee (2006) introduced the concept of *Proof* to describe for software agents (or human users) why they should believe a retrieved result. Hendler et al. (2011) argues that Proof can be achieved, if Linked Data Principles are applied to the dataset itself as metadata, including information about authorship, currency, license etc. (Hendler et al., 2011). One mechanism for publishing this type of metadata is Semantic Sitemaps which are an extension of the well-established Sitemaps protocol. In practice, however, semantic metadata are processed by different loosely coupled systems which

makes tracking, propagating, and querying difficult (Jacques et al., 2012).

Finally, the concept of *Trust* is suggested to be attained through digital signatures which are envisioned to check if data really comes from the claimed and trusted source (Berners-Lee, 2006). This, however, has not yet progressed far beyond a vision.

As discussed above, current research envisions how trustworthiness can be achieved, and the framework for different systems and technologies have been theoretically outlined. However, practical solutions have not yet been developed, thus it is not possible to state that the concept of SW is fully trustworthy.

#### 4 METHOD AND ANALYSIS

The study attempts to identify issues within the existing practice of Danish news journalism which might potentially be improved by SW technologies. This is examined qualitatively through a series of interviews with six news journalists and two digital editors from four of the largest Danish news media organisations. The same group of journalists and editors have been invited to engage in multiple Participatory Design (PD) activities. As a subgenre of qualitative research, PD aims not only at describing the social world, but also at contributing to the improvement of it by inclusion of stakeholders in the development of new services (Brandt, Binder, & Sanders, 2013). In this study, activities and techniques such as *The Future Workshop*, *Scenarios*, *The Magic If*, and *Prototyping* (Brandt et al., 2013) has been applied to explore: *How SW technologies can improve the journalistic practice?* And: *How the user experience of Danish news journalism can be improved in the context of SW?*

Transcripts of the qualitative interviews and PD studies have been categorised inductively after a grounded theory approach and analysed in the context of SW. Qualitative analysis reveals three areas within Danish news journalism with significant potential of improvement in the context of SW.

#### 4.1 Semantic Archive of Sources and Contact Details

The first area concerns the challenge of finding the right person to comment on a specific topic. Journalists often research in related news articles to find relevant sources. Such searches are however difficult to perform as persons are not semantically related to key terms or organisations. This part of the journalistic work process can be improved by implementing a semantic database of sources and contact details allowing a media’s journalists and editors to quickly identify relevant sources and their contacts in relation to specific search terms. The analysis finds that information about all persons described in previously published articles need to be annotated to form a database of already used sources. This application is referred to as Semantic archive of sources and contact details. Figure 2 illustrates the simplest graph needed to build a functional archive of sources and contact details, where journalists can easily evaluate a source’s area of expertise (*worksFor*, *knowsAbout*, *areaOfExpertise*) and reliability in a specific context (*jobTitle*, *qualifications*). The graph also meets the third Linked Data Principle about triples that should be included in a resource’s RDF/XML description. All resources (represented by circles in Figure 2) must be organised in a formalised ontology with class hierarchies and domain-/range- relations. It is recommended to apply existing standard vocabularies as resource- and

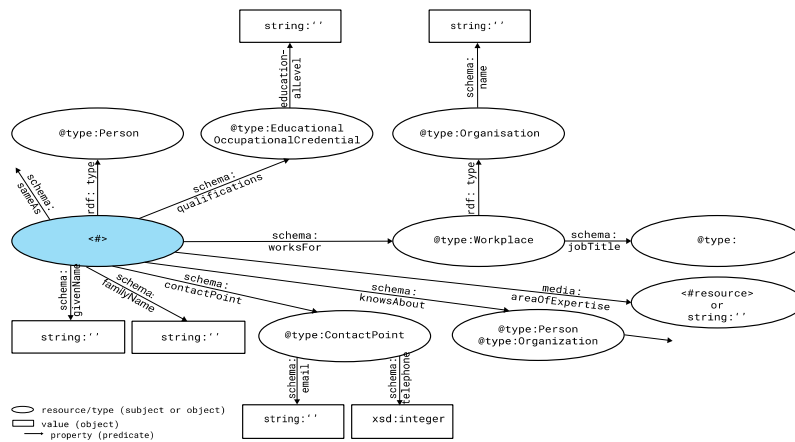


Figure 2: Minimum graph structure for semantic description of sources and contact details.

property- URIs can then be reused together with the underlying ontology of the standard vocabulary (Berners-Lee, 2006). In the example illustrated in Figure 2, the standard vocabulary Schema.org<sup>2</sup> is applied. Most standard vocabularies include resources phrased only in English which might entail a language issue when used for annotating news articles in Danish. This language differentiation presents a weakness of semantic annotation as it can be hard to integrate and might cause inconsistency or missing links. Thus, it is recommended to create new resources in Danish in a media’s controlled vocabulary and link these to equivalent English resources – if possible, in existing standard vocabularies – via *owl:sameAs*-relations. This ensures that any *Danish* resource becomes part of a global graph and allows standardised software to be applied.

Semantic annotation can be implemented in each news article using RDFa. This is however a very time-consuming process and requires basic knowledge of HTML and the RDFa syntax which journalists are not expected to possess. Thus, it is necessary to develop and implement systems that can automatically or semi-automatically write the annotation. In 2008, Thomson Reuters launched the linked data entity extractor Calais which is a Web service capable of annotating documents with URIs of places, people, and organisations mentioned in unstructured text such as news articles (Hendler, Heath, & Bizer, 2011). Today, Calais is integrated in the latest version of Drupal which is the content management system (CMS) used by multiple Danish news media organisations.

To reach a critical number of annotated concepts, the established SW community<sup>3</sup> emphasises large-scale cooperation and user contributions. Same approach needs to be applied within news media organisations: Every journalist must contribute and add pieces of semantic information when she comes across new or updated data. Thus, this type of application is not suitable for smaller organisations. In fact, it is recommended that multiple Danish news media organisations join forces and build a database of sources and contact details together.

Once semantic annotation is applied, it is possible to query sources and their contact details related to specific organisations, persons, or topics using SPARQL-queries. Journalists are not trained in constructing such queries, and it is recommended to

develop a search panel and design a user interface to guide the construction and formatting of these. Figure 3 illustrates how a search panel can be constructed. The search panel suggests three fixed search parameters (the properties: *areaOfExpertise*, *worksFor*, and *qualifications*) to be matched with different search terms (values). It is possible to type each search term and choose between suggested predefined terms from a media’s controlled vocabulary. These appear from a *drop-down list* as the user starts typing. The user interface ensures that the combination of properties and values are matched and formatted correctly following the SPARQL syntax. The technical implementation of this formatting can be performed using the library SPARQL Lib. This type of application raises questions about privacy and GDPR regulations which should be examined and discussed separately.

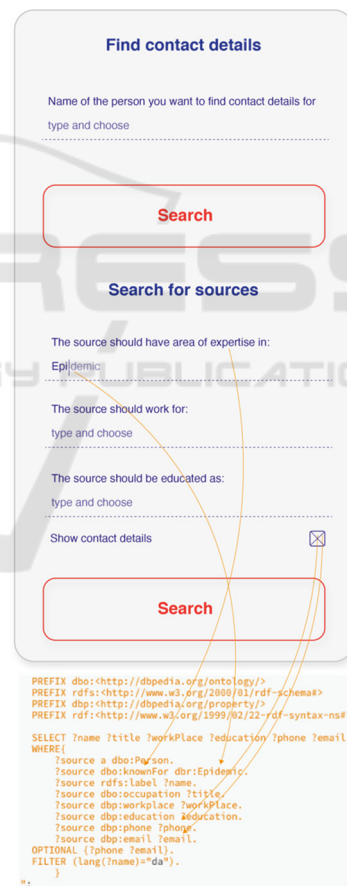


Figure 3: Search panel for sources and contact details.

<sup>2</sup> Schema.org was launched in 2011 by Google, Yahoo, and Bing as a standard for semantic mark-up of web pages (Bradley, 2013). Since then, Schema.org has grown to become one of the most popular standard vocabularies.

<sup>3</sup> The Linked Open Data Cloud (<https://lod-cloud.net/#>) keeps track of how many datasets have been published in the linked data format. The organisation also contributes to conferences and advisory boards.

## 4.2 Internal Semantic News Article Search

The second area of interest concerns issues of finding previously published articles related to a specific concept. When news break, journalists often search for context, and sources in related news articles published by their own media or by others. Journalists and editors however experience inconsistency and limitations in standard search engines such as Google, e.g. it can be difficult to search for articles published a long time ago. This process can be empowered by semantic annotation: If all persons, organisations, places, and key terms described in a media's archive of articles are semantically annotated, it is possible to perform thorough and complex search queries. This application is referred to as Internal Semantic News Article Search. This type of application requires concepts to be annotated as RDF triples the same way as for the semantic archive of sources and contact details. This annotation is however much more comprehensive and requires use of multiple vocabularies to describe relevant relations between different domains. Furthermore, it is important to distinguish real-world objects from the HTML document (the news article) that describe these, and it is recommended to annotate not only concepts and domains but also metadata about the Web document and its relation to the content (Hendler et al., 2011). The rNews ontology<sup>4</sup> can be applied to annotate information about author, publication data, and thumbnail-URL. As an example, this allows users of the semantic news article search application to search for concepts described by a specific journalist or written within a specific timeframe.

In order for journalists to apply semantic RDFa annotation as they write and publish news articles, a robust annotation system – preferably integrated into the CMS – is needed. However, even with an annotation system such as N.Y.T.'s Editor<sup>5</sup>, the process of semantically describing an entire news article requires more time, than journalists spend on tagging articles today. Also, current technology still requires manual editing to secure quality mark-up, and it should be considered whether this is a task for journalists, editors, or maybe even dedicated mark-up specialist to do. These issues are highly relevant and

worth a deeper study with journalists and other stakeholders.

Once all concepts are semantically annotated, it is possible to develop an application to perform complex and precise searches within a media's database of linked data. In fact, such an application already exists: N.Y.T. has created a semantic search application<sup>6</sup> for their archive of articles – consisting of news articles from 1981 N.Y.T. Developer (n.d.). The application is designed as an API for public use over the HTTP and is used as a case study in this exploration.

Apart from supporting the research phase, some participants suggest that an internal semantic news article search application can be used for additional purposes: The function can be used to insert links to related articles in a specific news article. This is something Danish journalists currently do manually because existing technologies are not good enough. An automation of this process saves time and enables new links – which are not present to the journalist – to be added. Participants also suggest that internal semantic news article search can be used to analyse how minorities are represented and used as sources in the overall news coverage. This type of analysis can be applied as a tool for the media to shed light on biases or specific discourses, but it can also provide research for new articles worth reading for the public.

The sections above describe how a semantic search application can be implemented for a media's own archive of articles. Journalists however often research in other media's archives too, and it would increase usability remarkably, if not only one Danish news media organisations, but all news organisations world-wide decided to semantically annotate their articles and provide open API search tools.

## 4.3 Semantic Infobox: Summary

The third area of interest targets improvement of user experiences and concerns the issue of adding encyclopaedic information in a short amount of time. The analysis finds that Danish news articles do not deploy online opportunities of providing additional information, and that Danish news journalists are not willing to spend more time than absolutely necessary on writing additional, encyclopaedic information

<sup>4</sup> The vocabulary is designed by the International Press Telecommunications Council (IPTC) to ensure consistency in how news media annotate metadata.

<sup>5</sup> In 2015, New York Times launched *Editor* which a semi-automated tool for annotating news articles with semantic information. The tool comprises a simple text

editor, supported by a set of networked microservices that are trained to apply specialised N.Y.T. resources to text documents.

<sup>6</sup> <https://developer.nytimes.com/docs/semantic-api-product/1/overview>

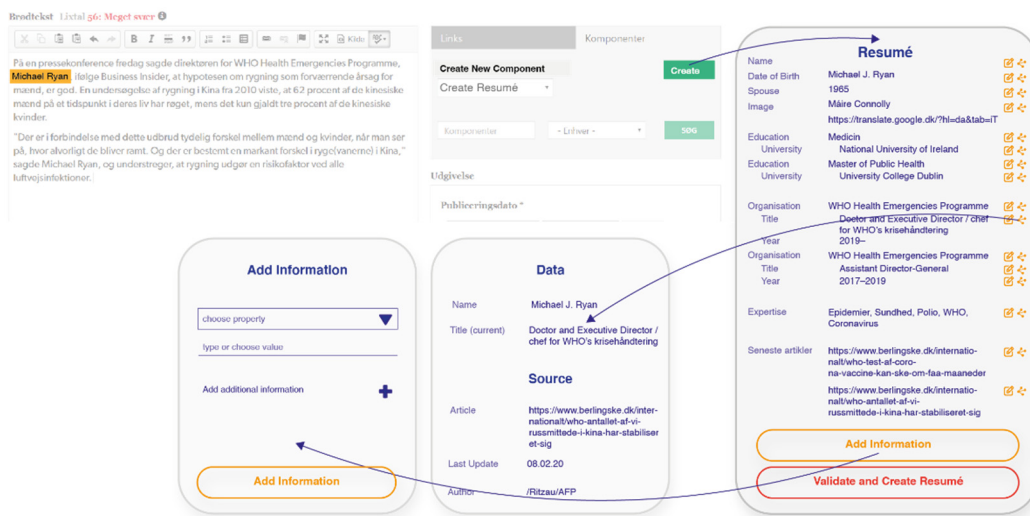


Figure 4: Interface for adding autogenerated summaries.

about persons or concepts mentioned in news articles. On the other hand, it can be concluded that such additional information contributes to transparency and ultimately supports a media's trustworthiness.

To ensure the quality of information, it is recommended to focus on one type of infobox: Autogenerated summaries – and as a starting point design the application to rely exclusively on information from a media's own database of annotated information. This application is referred to as Semantic Infobox: Summary.

The analysis finds that autogenerated summaries must include information about a person's name, job title, workplace, educational background, and seniority. This way integrated summaries can strengthen a media's trustworthiness as they document why professors or other authorities are chosen as expert sources. Additionally, summaries should include personal information such as date of birth, family relations, and information about previous jobs and memberships. This allows readers of the news article to easily recap information about the person. Ultimately, news journalists can also use the application as a research tool. For this type of application, it is important that all string values are phrased in Danish, and that all resources describing anything else than names are linked to resources phrased in Danish via *owl:sameAs*-relations. Similarly, all properties need to be linked to equivalent properties in Danish. This allows the autogenerated summary to be displayed with both properties and values phrased in Danish.

An application for autogenerated summaries differs from the two previous solutions as it not only support research but contributes with information

displayed directly as part of the news article. This highly increases requirements for reliable information as one incorrect information contained in an infobox might affect the trustworthiness of all other content on the media's platform. Possibilities for fact-checking can be provided as metadata such as when the summary was last updated, and what sources the data is collected from. This kind of transparency and traceability can be implemented using Semantic Sitemaps which allows metadata to be included in the RDF/XML descriptions.

Figure 4 demonstrates how a user interface – integrated in the CMS – for implementing and fact-checking autogenerated summaries can be designed. The journalist simply highlights a person's name and uses the middle panel to create a summary. A display then appears showing all relevant information about that person. For each line of information, the journalist is provided with options to edit data or inspect the data source. This way, the journalist can review when a person's current job title was last updated or delete outdated information. It is also possible for the journalist to add new information using predefined properties and predefined or typed values. Technically, summaries can be queried and formatted using SPARQL-queries and SPARQL Lib the same way as results are retrieved and displayed for the semantic archive of sources and contact detail application.

Autogenerated summaries as described in the sections above rely exclusively on linked data from a media's own archive of news articles. This guarantees journalists and users that information contained in the summary at some stage has been fact-checked and edited by a journalist, but it also means that the application can only generate summaries for persons

previously mentioned by the media. In order to truly support news journalists, and to truly take advantage of the concept of SW the application should be extended to rely also on external sources of linked data. It is recommended to develop some kind of certification to guarantee that the datasets live up to current GDPR regulations and are maintained and up to date, e.g. guaranteeing that the dataset is reviewed at least once a month, and that a person registered with name and contact details is responsible for this. Establishing national or even international standards for the quality and maintenance of linked data allows media organisations – and other organisations – to share and reuse encyclopaedic data from each other. The analysis demonstrates that news journalists across Danish news media organisations work in very similar ways and with extended focus on research and fact-checking. These conventions might ease the process of defining a set of standards for summary data, but legal and practical implementation of such certification requires further research.

## 5 DISCUSSION AND CONCLUSIONS

The core intention of SW is to make content machine-readable in order to improve findability and enable knowledge- and context-based information to be generated. This way semantic annotation can be seen as pure preparation for AI agents. However, this study demonstrates that AI in the context of Danish news media still seems distant. Remarkably little research explores aspects of reliability and objectivity within the field of SW even though major challenges on how to secure trustworthy information need to be solved. As demonstrated in this study, these challenges need to be solved not only locally, but as standards or certifications agreed upon by all trustworthy linked data providers. Another area which needs attention in this context is the concept of *neutral ontologies* (Uschold and Gruninger, 2004). The concept of ontologies presupposes that the entire world can be objectively categorised. However, several current debates illustrate how this – in reality – is often negotiable or political dependent. Thus, those in charge of annotating information and constructing underlying ontologies can easily influence the way we understand our surroundings which has proved to be extremely powerful. In contrast to the power of current media, this impact is much less visible and harder to trace and should be an area of extreme interest for researchers in the field of SW and in the

field of journalism. In the context of Danish news media, this paper has, in a practical manner, explored potentials of SW technologies exposing three areas where these technologies can be used to make significant improvement. Before this can be realised, basic research on implementation of annotation tools and standards for trustworthiness is still needed.

## REFERENCES

- Alam, B., Birbeck, M., McCarron, S., Herman, I. (2015). Syntax and processing rules for embedding RDF through attributes.
- Berners-Lee, T. (2006). Linked Data. Retrieved from: <http://www.w3.org/Designissues/LinkedData.html>
- Brandt, E., Binder, T., Sanders, E.B.N. (2013). Tools and techniques: Ways to engage telling, making and enacting. In *Routledge International Handbook of Participatory Design* (p.145–182). Routledge.
- Domingue, J., Fensel, D., Hendler, J. (2011). *Handbook of Semantic Web Technologies*. Berlin: Springer, vol. 1.
- Goddard, L., Byrne, G. (2010). Linked Data tools: Semantic Web for the masses. In *First Monday*, 15(11).
- Hendler, J., Heath, T., Bizer, C. (2011). *Synthesis Lectures on the Semantic Web: Theory and Technology. Linked Data evolving the Web into a Global Data Space*. Morgan & Claypool Publishers.
- Heravi, B., McGinnis, J. (n.d.). A Framework for Social Semantic Journalism.
- Jacques, Y., Anibaldi, S., Celli, F., Subirats, I., Stellato, A., Keizer, J. (2012). Proof and Trust in the OpenAGRI Implementation. *Int'l Conf. on Dublin Core and Metadata Applications*.
- Miller, P. (2008). Sir Tim Berners-Lee: Semantic Web is open for business. Retrieved from: <http://blogs.zdnet.com/semantic-web/>
- N.Y.T. Developer (n.d.). Semantic API. Retrieved from: <https://developer.nytimes.com/docs/semantic-api-product/1/overview>
- N.Y.T Labs (2015). Editor (2015). Retrieved from: <https://nytlabs.com/projects/editor.html>
- Pandey, R., Sanjay, D. (2010). Interoperability between Semantic Web Layers: A Communicating Agent Approach. In *International Journal of Computer Application*, 12(3).
- Raimond, Y., Scott, T., Oliver, S., Sinclair, P., & Smethurst, M. (n.d.). Use of Semantic Web technologies on the BBC Web Sites.
- Rudnik, C., Ehrhart, T., Ferret, O., Teysou, D., Troncy, R., Tannier, X. (2019). Searching News Articles Using an Event Knowledge Graph Leveraged by Wikidata. *2019 IW3C2*.
- Uschold, M., & Gruninger, M. (2004). Ontologies and semantics for seamless connectivity. In *SIGMOD*, 33(4), pp. 58-64.
- Wood, D., Marsha, Z., Luke, R., Hausenblas, M. (2013). *Structured Data on the Web*. Manning Publications.