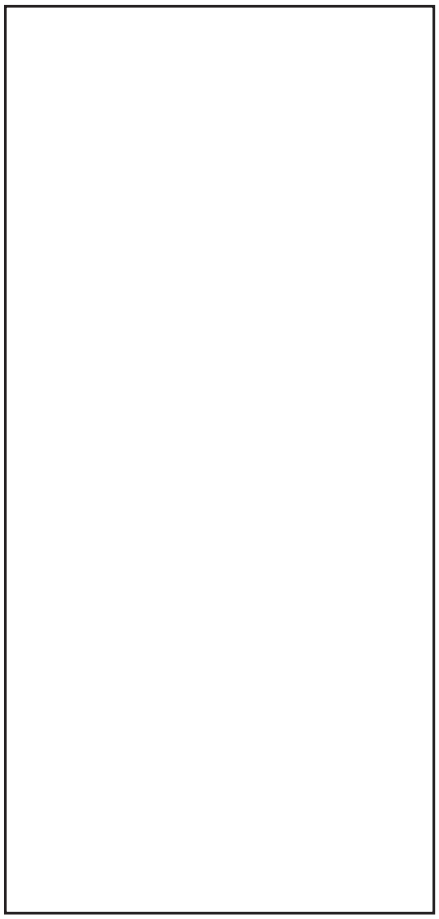
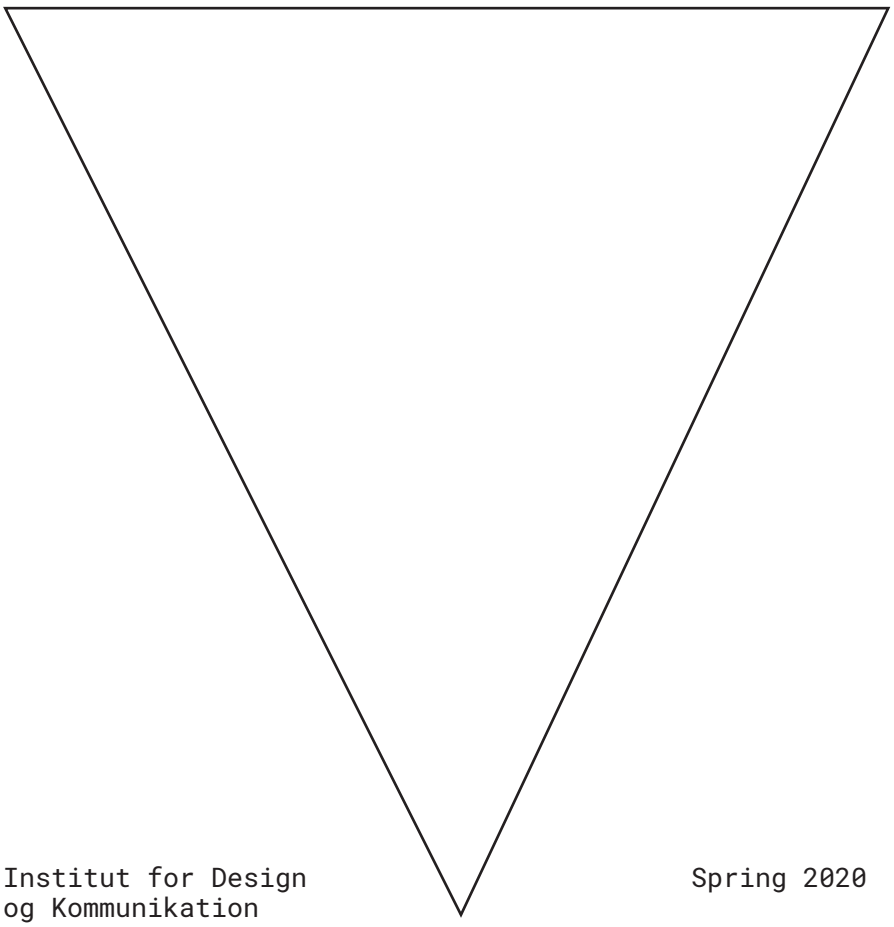
Astrid Ildor
MSc in Information Technology

Danish News Journalism in the Context of Semantic Web

Danish Title: Dansk nyhedsjournalistik på det Semantiske Web





Syddansk Universitet
Thesis project

Institut for Design og Kommunikation

Spring 2020

Abstract

For more than a decade, Semantic Web has been considered open for business (Miller, 2008). Semantic Web is best understood as an extension of the existing Web which contains immense potential to transform the Web into a distributed reasoning machine that not only can execute extremely precise searches but also analyse existing data and create new knowledge (Berners-Lee, Hendler, & Lassila, 2001).

In the same period of time, media organisations have implemented comprehensive online practices: News stories break and are published on the Web, and desk research is an essential tool to any news journalist (Veglis & Bratsas, 2017).

However, only few news media organisations have made efforts to take advantage of Semantic Web technologies (Domingue, Fensel, & Hendler, 2011, p. 57), and remarkably little research has systematically examined insights of editors and journalists to better adopt and deploy these evolving technologies.

To fill the gap in literature and to motivate further innovation, this study explores how existing work processes of Danish news journalists and the user experience of Danish news journalism can be improved in the context of Semantic Web?

The examination draws on qualitative interviews and participatory design studies with six writing news journalists and two digital editors from four of the largest Danish news media organisations. Areas of interest are analysed and technically discussed in the context of Semantic Web.

This inductive approach and combination of domain-specific insights and technical analysis is new to both developers in the news media industry and scholars within the field of Semantic Web research.

The examination reveals three areas with significant potential of improvement. The first area concerns journalists' challenge of finding the right person to comment on or evaluate a specific topic. The second area concerns issues of finding previously published articles related to a specific concept. Finally, the third area targets the need for generating additional encyclopaedic infoboxes in a short amount of time.
Each area is translated into and examined as a type of Semantic Web application:

- Application I: Semantic archive of sources and contact details
- Application II: Internal semantic news article search
- Application III: Semantic infobox: Summary

It is demonstrated how linked data[1] can be annotated[2] and queried[3] for the proposed applications.

The study concludes that Application II + III require profound annotation of all persons, places, organisations, and key terms mentioned in a media's archive of news articles. Even with support from existing annotation software, this process is highly time consuming and requires implementation of completely new work processes. Application I requires less thorough annotation and is thereby a more realistic suggestion for a Semantic Web application that can benefit Danish news media in near future.

Another major challenge is the issue of securing trustworthiness and proof of documentation which is especially urgent when applications are designed to rely on external sources. Basic research on how trustworthy information can be guaranteed in the context of Semantic Web is still needed before theory can be turned into practice and benefit Danish news media.

Keywords
Semantic Web · Linked Data · Web 3.0 · News Media · Journalism · RDF · Semantic Annotation · URI · Web Ontologies · Application Development

1 Linked data can be understood as semantically annotated information. Since 2009, the term linked data has been considered a means to reach a Semantic Web (Hendler, Heath, & Bizer, 2009., p. 17) This is illustrated in the model Semantic Web Stack (see Fig.06).

2 Annotation or semantic mark-up is the process of attaching additional information to various concepts in a given Web document. Unlike classic text annotation, semantic annotation are used by machines to refer to and can be applied using standard vocabularies or ontologies (Kyrnin, 2020) – see also glossary list in Appendix 9.22.

3 A query is a form of questioning, in a line of inquiry (Cambridge Dictionary: Query). When querying information on Semantic Web, the query language SPARQL needs to be applied – see also Sec.3.5 and the glossary list in Appendix 9.22.

Table of Contents

List of figures

List of tables

List of code snippets

1.0 Introduction

Chapter 1 provides an overview of the subject of this study: News journalism in the context of Semantic Web.
Sec.1.1 summarises the problem background and positioning of research. Research questions are then presented in Sec.1.2, before Sec.1.3 reasons for delimitations of the study. Remarks on terminology can be found in Sec.1.4. Finally, Sec.1.5 outlines the structure of this work.

Today, most news media organisations publish information on the Web with a publish-and-forget mindset (Goddard, Lisa & Byrne, 2010): Once a news article is published, the information it contains devalue as the document gets discarded in the messy place of the unstructured Web. Anyone who has tried to catch up on news event, knows how hard it is to get an overview of old news – even journalists find it difficult to use their own coverage for structured research (see Sec.4.3).
In a time where information and big data is among the most valuable resources, it is a paradox that news media organisations who collect and publish trustworthy information on an hourly basis do not pay attention towards the potential of archiving and reusing that information.

The concept of Semantic Web contains immense potential to transform the Web into a distributed reasoning machine that not only can execute extremely precise searches but also analyse existing data and create new knowledge (Goddard, Lisa & Byrne, 2010). This can potentially improve the process of any news journalist and entail innovative storytelling and knowledge mediation.

However, new technology does not arrive on its own. Before media organisations can be expected to make large investments in archives of linked data, publis-

hers, editors, and journalists must be convinced that there are costly problems associated with their current suite of technologies, and that Semantic Web applications can solve these and provide good return on their investments.
On the other hand, news media has long constituted an area of interest for Semantic Web researchers, but remarkable little research combines the knowledge of technologists with insights of editors and journalists.

To fill the gap in literature, this study intends to examine how the work process of Danish news journalists and the user experience of Danish news journalism can be improved in the context of Semantic Web?
First qualitative research is applied to examine Danish news journalism practice, then areas of interest are analysed and technically discussed in the context of Semantic Web. The aim is to come up with a set of recommendations on where to focus future innovation and research.

1.1 Positioning of research

This examination is executed in the cross section of Semantic Web research and journalism research. Before research questions and -design are introduced, this section outlines the problem background and positioning of research which is based on a small literature review. The review is conducted using the search methodology block search[1] **in relevant academic databases (see Appendix 9.1).**

1 The method combines different keywords to a search profile which is described in detail in Appendix 9.1.

1.1.1 Positioning of research in the field of Semantic Web research

As commonly known, the Web was envisioned in 1989 by Berners-Lee to tackle the problem of knowledge sharing between co-workers at CERN where he worked at the time. The immense power of the Web lies in its ability to link one document to another through hyperlinks. This interconnectivity enables users to easily access information and browse additional sources (Berners-Lee, Hendler, & Lassila, 2001). The concept is however not impeccable, and in 1994 the same Berners-Lee articulated the vision of a Semantic Web to solve two specific problems central to the Web (Domingue, Fensel, & Hendler, 2011, p. 9). Semantic Web can this way be characterised as an extension of the existing Web.

The first of the aforementioned problems concerns accessing data: String-based matching algorithms were – and to a large extend still are – used to retrieve documents according to a given search query. This creates problems for ambiguous terms, e.g. the term Denmark is a country, a kingdom, and a city in the state of South Carolina. Moreover, complex matching is not possible with current search engines: Often, the answer to a query exists on the Web but requires integration of multiple sources which is not possible with standard search engines (see Sec.3.1 for specific examples of this). Finally, a significant number of websites are generated from databases, but the underlying data of these are hidden behind the presented HTML which hinders reusability (Domingue et al., 2011, p. 10).

The second problem concerns delegation: When users browse the Web, their computers act simply as rendering devices displaying text, audio, and video content. All inference, computation, and delegating tasks such as the integra-

tion of information, data analysis, and sensemaking are left to the user (Domingue et al., 2011, p. 10).

In order to solve these problems, not just documents but all data points constituting the Web need to be interconnected. This can be achieved by semantically annotating data and is referred to as linked data which constitutes the building blocks of Semantic Web (Berners-Lee et al., 2001).

A large amount of Semantic Web research (Domingue et al., 2011; Feitosa, Dermeval, Farias Lóscio, & Isotani, 2017; Hendler, Heath, & Bizer, 2011) has been devoted to the issue of constructing, publishing, and querying linked data. Quantitative and computational studies have contributed with theoretical models, standards, and principles to an extent where Semantic Web can be considered a theory of its own.
This point is supported by Semantic Web professors Tom Heath and Christian Bizer who in 2011 published the book Linked Data – Evolving the Web into a Global Data Space with the subtitle Synthesis lectures on the Semantic Web: Theory and Technology. Furthermore, the concept of Semantic Web can be distinguished from traditional information systems as it constitutes a set of standards and generic needs which can be realised in different components or applications (Domingue, Fensel, & Hendler, 2011, p. 49).

In February 2008, annotation tools reached a point of development where Berners-Lee and the World Wide Web Consortium (W3C) declared Semantic Web open for business (Miller, 2008).

Coinciding, multiple news media organisations launched initiatives for annotating their archives of news articles in order to generate Semantic Web applications (see Fig.01 below).

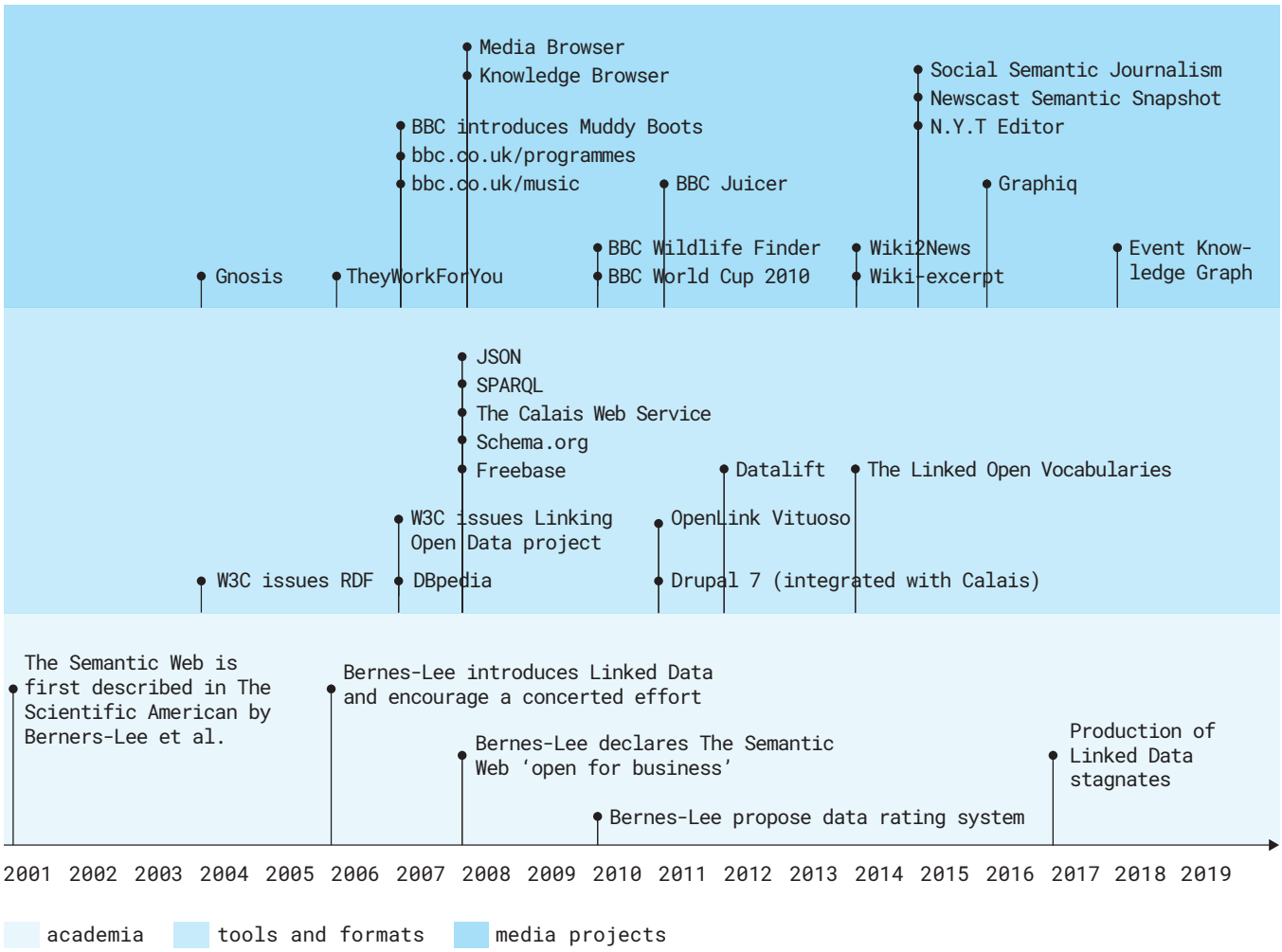


Fig.01 Semantic Web – timeline. See Appendix 9.2 for more detailed description of media projects

In 2008, Thomson Reuters launched a Web service capable of extracting entities and relationships in text documents such as news articles and annotating these with linked data URIs (see Sec.3.2). According to Hendler et al. (2011):

> (...) such services bridge linked data and conventional hypertext documents, potentially allowing documents such as blog posts or news articles to be enhanced with relevant pictures or background data.
>
> **(Hendler et al., 2011, p. 35)**

In cooperation with Rattle Research, the British Broadcasting Company (BBC) has developed a similar service to extract main entities from BBC news articles

and match them with resources in DBpedia[2] (Kobilarov et al., 2009, p. 727). This annotation has empowered applications such as BBC Wildlife Finder which re-purposes data from different sources including Wikipedia, WWF, and the IUCN's Red List of Threated Species and combines it with natural world footage from the BBC archive (Raimond, Scott, Oliver, Sinclair, & Smethurst, n.d.). This way users can navigate additional information about the featured wildlife.

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). This has allowed N.Y.T. to annotate their archive of articles from 1981 to today (N.Y.T. Developer, n.d.) and on this basis generate multiple Semantic Web applications (see Appendix 9.2).

A handful other smaller news media organisations and researchers have launched similar Semantic Web applications – a collection of these can be found in Appendix 9.2 and is further discussed in Sec.6.1. Review of the collection demonstrates that none of these applications seem to be based on systematic empirical insights of news journalists or publishers.

It can be concluded that Semantic Web researchers have established a theoretical framework to support linked data production. W3C Semantic Web standards have been mature for several years, and real-world tools are available for publishing linked data. However, only few news media organisations have made efforts to adopt Semantic Web technologies (Goddard, Lisa & Byrne, 2010). This hesitant realisation has resulted in several critics doubting that Semantic Web will be anything but a high-tech fantasy (Finlayson, 2010) which is supported by that fact that the Linked Open Data Cloud who monitor the global production of linked data has reported remarkable stagnation since 2017.

This study aims to motivate further development and investment by demon-

2 The open, linked database DBpedia is introduced in Sec.3.3.1 and in the glossary list in Appendix 9.22.

strating how Semantic Web technologies can solve specific current issues within Danish news media. The examination contributes to the field of Semantic Web research by systematically bringing forward insights of Danish news journalists and editors. This qualitative and inductive approach (see Sec.2.1–2.2) is new to both developers within the news media industry and researchers within the field of Semantic Web research.
Semantic Web – viewed as a theory or not – emphasises the importance of indexing and archiving information (see Chap.3.0) which today seems neglected in Danish news media. Thus, analysis applied on this sector is expected to reveal valuable insights.

1.1.2 Positioning of research in the field of journalism research

This section outlines how the field of journalism research examines concepts of data journalism and Semantic Web technologies. This is relevant for the positioning of this study; however, the section should not be read as a full literature review of journalism research.

When searching for the terms Semantic Web, linked data, or Web 3.0 in SAGE Publications' Encyclopaedia of Journalism null results are returned which indicates that Semantic Web is not yet an established term in the field of journalism research.
Review of relevant databases (see Appendix 9.1), however, demonstrates that few scholars (Creamer, 2008; Finlayson, 2010; Veglis & Bratsas, 2017) have examined the concept and acknowledge that Semantic Web – also known as Web 3.0 (see Sec.1.4) – should be of great interest for the news media industry:

> If Web 1.0 was about old-media companies making half-hearted gestures at that online thing, and Web 2.0 was a brisk reminder that (...) the internet

> is a very real and open thing where walls and control don't work well, then part of Web 3.0 will be about figuring out how to monetise that openness. **(Creamer, 2008)**

Veglis & Bratsas (2017) define data journalism as a journalism speciality in which numerical data are used in the production and distribution of information**. The authors relate this type of journalism to Semantic Web and argues that it has evolved in recent years** as access to digital data has remarkably increased **(Veglis & Bratsas, 2017, p.235).**

In data journalism, journalists are required to master skills of seeking information on the Web, visualising data with the help of various applications, and publishing material on the Web (Veglis & Bratsas, 2017, p.236). To this work practice, the authors add what they call Web 3.0 skills **which include SPARQL Protocol and RDF Query Language (SPARQL) queries (see Sec.3.5) which they recommend journalism schools start teaching (Veglis & Bratsas, 2017, p. 236).**

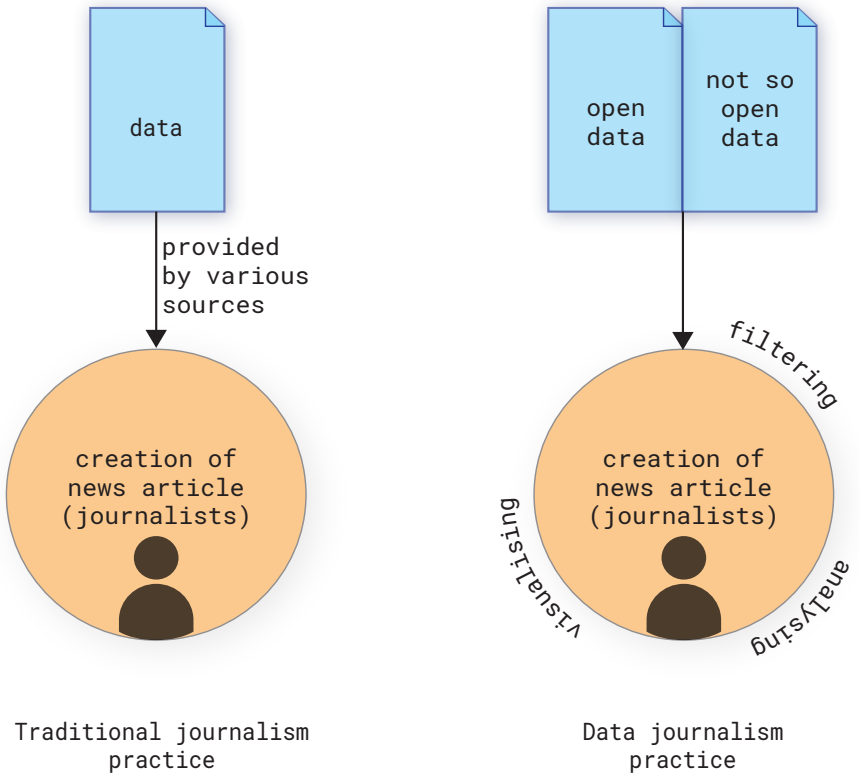


Fig.02 Traditional journalism versus data journalism (after Veglis & Bratsas, 2017)

Finlayson (2010) examines different thought-up examples of Semantic Web technologies in the context of news journalism. Among others, he focuses on the launch of open government data in the US and UK which he describes as a treasure trove that journalists can analyse with the potential to inspire thousands of new stories (Finlayson, 2010, p. 62).
He concludes that Semantic Web technologies can potentially solve several problems for the news media industry. Journalists and editors however do not acknowledge these as problems, thus development cannot be expected to happen organically:

> Many people, particularly those who would have to pay their business apply Semantic Web concepts do not see the immediate benefit of such an effort. (...) Getting everyone to agree and then act on one way to describe data is going to happen only when there are compelling economic reasons, or it becomes so easy to do that there is no reason not to.
> (Finlayson, 2010, p. 63)

This study contributes to the field of journalism research by identifying current issues within Danish news media and analysing these in a context – the concept of Semantic Web – which has not been examined before.

1.2 Research question

The aim of this study is to bridge technical knowledge and domain-specific insights of news journalists in order to develop meaningful Semantic Web application for Danish news media. Thus, the problem statement of this work is:

How can the work process of Danish news journalists and the user experience of Danish news journalism be improved in the context of Semantic Web?

- **What areas of Danish news journalism practice can be improved by Semantic Web technologies and how might this be done? (RQ1)**

- **How can the user experience of Danish news journalism be improved by Semantic Web technologies? (RQ2)**
- **What usability considerations and technical requirements can be found to realise solutions proposed as answers to RQ1 and RQ2? (RQ3)**

RQ1 addresses a qualitative examination of the work process of Danish news journalists and seek to identify areas where Semantic Web technologies might contribute to a more powerful practice. The aim is deduce recommendations on where to focus future innovation and specific proposals of Semantic Web applications to solve the identified issues.

RQ2 concerns how online news articles are perceived. This part of the study qualitatively examines where the user experience of Danish news articles is insufficient and how Semantic Web can contribute to improve it. The aim is to propose specific examples of Semantic Web applications to improve aspects of user experience.

Finally, RQ3 addresses a more technical examination of Semantic Web applications proposed as answers to RQ1 and RQ2. This part of the study examines development and implementation of three types of Semantic Web applications in the context of Danish news media. The aim is to discuss usability considerations, technical challenges and possibilities on a more detailed level rather than to provide step-by-step descriptions on how the applications can be implemented. Detailed description of the research design can be found in Sec.2.3.

1.3 Delimitations

The scope of this study is limited, thus delimitations and constraints are needed.

The exploration encompasses written news coverage, meaning textual articles communicating factual information. Potentials and possibilities of audio-visual material in the context of Semantic Web is not included as Semantic Web technologies for this type of data is not yet nearly as developed as for written information.

This study aims at describing how Semantic Web technologies can contribute to Danish news media in near future. The exploration is grounded in existing journalism practice and technologies and the limits of these anno 2020. The concept of Semantic Web includes visions of the entire Web as one interconnected global graph (see Sec.3.3). This is, however, not likely to be realised within the next ten years (Domingue et al., 2011, p.585), hence such scenarios are not included in the scope of this study.

Challenges of ethical and legal character is touched upon throughout the exploration, but thorough discussions on data ownership, privacy, GDRP, profiling, and data mining are deliberately left out as the extent of these topics requires seperate examinations. Further research on these higly important topics are however strongly recommended (see Sec.7.1).

This study bridges qualitative analysis and technical examination as the combination of these approaches are expected to reveal valuable insights. This scope however limits the level of technical detail, and phases of implementation, integration and test runs are left out. The aim of this part of the examination is to illuminate and discuss domain-specific possibilities and challenges rather than to provide a complete recipe on how to build Semantic Web applications. Finally, limits and possibilities for generalisation are discussed in Sec.2.4.

1.4 Remarks on terminology

The terms Web 3.0, Semantic Web and linked data are often used to describe

the same concept. Semantic Web is the term Berners-Lee first used, when he introduced his vision in 1994. Later, the term Web 3.0 has been used in more popular contexts to put the vision in relation to the two former paradigms of the Web – Web 1.0 and Web 2.0 (see Appendix 9.22).
In this study, Semantic Web is used to describe the overall concept of an interconnected Web of data.

Since 2009, the term linked data has been considered a means to reach a Semantic Web (Hendler, Heath, & Bizer, 2009., p. 17) which is illustrated in the model Semantic Web Stack (see Fig.06). Technologies from the bottom of the stack up to the layer of Ontologies and Rules constitute what is known as linked data. These technologies are currently standardised and accepted by the W3C to build Semantic Web applications.
In this study, Semantic Web technologies is used as a general term to describe all of the technologies included in the Semantic Web Stack.

RQ2 examines user experience of Danish news articles. The term user experience is broadly used to evaluate services within Human-Computer Interaction (HCI) (Rogers, Preece, & Sharp, 2015, chap.1.6). Methods and techniques to measure user experience is widely discussed, and HCI researcher do not agree on a fixed definition of the term. It is out of the scope of this study to quantitatively measure aspects of user experience, instead the ISO-definition below is used to deploy an area of interest:

> User experience includes all the user's emotions, belief, preferences, perceptions, physical, and psychological responses, behaviour, and accomplishment.
> **(ISO 9241-210 (2010))**

The term user experience is applied in RQ2 instead of usability to include dimensions of satisfaction and the user's apprehension of a media trustworthiness.
A definition of usability can be found in Appendix 9.22 where additional key

words and technical terms applied throughout this study can be looked up and examined in brevity.

1.5 Structure of this study

The study at hand consists of nine chapters. After the introduction, Chapter 2.0 presents the research design and methodology including interview- and participatory design techniques.

The theoretical framework is then set out in Chapter 3.0.

Analysis and interpretation of the qualitative research is presented in Chapter 4.0 which leads up to Chapter 5.0 where key findings are translated into three Semantic Web applications which are then examined and discussed in technical detail.

The exploration concludes with comparison and outlook in Chapter 6.0 and a conclusion in Chapter 7.0.

A list of references can be found in Chapter 8.0 while appendices are collected in Chapter 9.0.

Each chapter is supported by figures and models. These are constructed as .svg-files, and the reader is invited to zoom in to grasp details and descriptions. For the same reason, it is recommended to read the study on screen.

2.0 Methodology

Chapter 2 explains the methodology applied in this study.
Sec.2.1 introduces qualitative research as approach and research strategy. Sec.2.2 reflects on philosophy of science and the paradigms employed in this study, before an overview of the research design is outlined in Sec.2.3. Applied principles for quality data collection are described in Sec.2.4. Finally, the method of qualitative interview and the methods and techniques of participatory design employed in this study is described in detail in Sec.2.5 and 2.6.

2.1 A qualitative research approach and strategy

The potential of new technology is hard to study as the future is notorously impossible to know. In some ways, this study is comparable with asking how the hyperlink could contribute to news journalism back in 1994? Today – as we know the answer – it is hard to imagine research of the late 1990's describing even half of that. It is however possible to come up with qualified guesses in order to prepare the industry to invest in and develop new technologies. That is exactly what this study aims to do.

First of all, this study attempts to identify issues within the existing practice of Danish news journalism which might potentially be improved by Semantic Web technologies. Such issues cannot be quantitatively identified or measured but requires detailed observation and examination of Danish news journalists, their work practices, experiences, needs, and wishes. This approach can be characterised as qualitative research and is applied because of its

ability to provide thick descriptions[3] of a problem or topic from the perspective of the local population (Ritchie et al., 2003, p.1).

Qualitative research is often characterised by an inductive knowledge generation where hypotheses are generated from analysis of emperical data rather than stated at the outset (Silverman, 2011). This characteristic also applies to the study at hand which aims to formulate a set of recommendations based on key findings, rather than to strengthen or refute a theory or hypothesis.

As a subgenre of qualitative research, Participatory Design (PD) aims not only at describing the social world, but also towards contributing to the improvement of it by including users in the development of new products and services (Brandt, Binder, & Sanders, 2013). In this study, PD techniques are applied (see Sec.2.4 and 2.6) to identify issues in current work practices and to to include Danish news journalists and editors in the development of Semantic Web applications.

Key findings of these qualitative studies are object of technical analysis and discussion (see Sec.2.3).

2.2 Reflections on philosophy of science

Qualitative research studies belong to the methodologically field of social science. In very broad terms, social science has been shaped by two overarching ontological positions: Realism and idealism (Ritchie, Lewis, Nicholls, & Ormston, 2003, p.4). Realism is based on the idea that there is an external reality which exists independently of people's understanding of it. Idealism, on the other hand, asserts that reality is fundamentally mind-dependent: It is

3 See term definition in Appendix 9.22

only knowable through socially constructed meanings.

The approach of this work broadly falls within the philosophical school known as critical realism (Robson, 2002) or subtle realism (Blaikie, 2007) which is grounded in interpretivism (Bryman, 1988) and stresses the importance of interpretation.

Ontologically, this means that reality is seen as something that exists independently of those who observe it but is only accessible through perception and interpretation. As first described by Dilthey (1883), social research should examine lived experiences in order to reveal the connections between the social, cultural, and historical aspects of people's lives. This way, the critical importance of participants' own interpretations is recognised.
In this perspective, Semantic Web technologies cannot be judged as something ultimately good or bad, useful or useless. The technology might or might not work, but the quality and use of it depends on the interpretation of the people using it. These interpretations rely on previously lived experiences, habits, and assumptions and are important to consider when developing new systems and applications.

Epistemological debates are also central to social science. One view holds that knowledge is generated inductively through bottom-up processes where patterns derive from observation. In contrast, others view knowledge generation as a top-down deductive process, where logically derived hypotheses are tested against observations (Ritchie et al., 2003, p.6).
Grounded in an interpretivist frame, this study applies an inductive approach using evidence – such as analysis of interviews and PD studies – as the genesis of a conclusion. It is however important to note that data cannot be interpreted on completely blank paper: The questions asked, the categories employed, and the interpretations generated are influenced by assumptions derived from previous work.

This leads to another epistemological discussion which concerns the relationship between researcher and the researched. The researcher of this study believes that in a social world, people are affected by the process of being studied, and that the relationship between the researcher and social phenomenon is interactive. Complete objectivity is not possible, but neutrality in the collection, interpretation, and presentation of data is worth striving for to avoid obvious bias and to keep focus on understanding and examining views and experiences of the participants. Ritchie et al. (2003) describes this as empathic neutrality.

In the view of subtle realism, researchers must also be reflective of their influence on the research process and its participants. In this context, it is important to note, that the researcher of this study is an educated journalist herself. This might influence the qualitative research of the journalistic practice as it can be hard to examine one's own habitat (Spradley, 1979). It is important not be locked in one's own interpretation or to influence the empirical data with personal insights. On the other hand, the double role might contribute positively to the difficult task of creating a meaningful interview guide and valuable PD activities as the researcher possesses a preunderstanding of the participant's framework of knowledge and references.

2.3 Research design

This section describes the overall research design as it is outlined in Fig.03 below. Detailed descriptions and planning of the applied methods can be found in Sec.2.5–2.6.

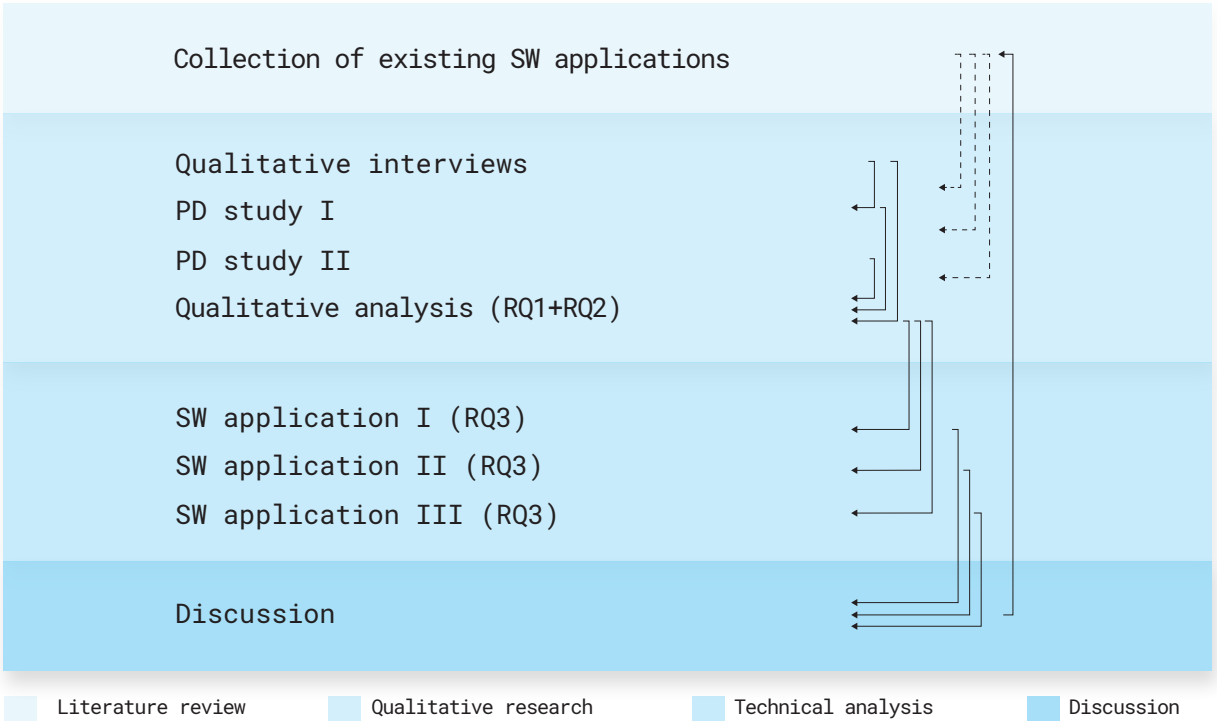


Fig.03 Research design

As a point of departure, the field of Semantic Web research and journalism research is examined through small scale literature reviews (see Sec.1.1 and Appendix 9.1). This examination includes a collection of existing Semantic Web applications developed for the media industry (see Appendix 9.2), and the examination serves as a knowledge pool for the design of a qualitative research study.

Qualitative research constitutes the core of this study and aims at examining the work process of Danish news journalists. This is done through qualitative interviews (see Sec.2.5) with Danish news journalists and editors. The aim of this series of interviews is two-folded: First, to provide empirical data – a detailed description of the journalistic practice – to analyse and discuss in the context of Semantic Web. Second, the interviews form a foundation for the first of two PD studies.
As a continuation of the qualitative interviews, a PD study with the same

group of journalists and editors is carried out. Participants are invited to engage in multiple PD activities (see Sec.2.6 and Appendix 9.4) which explore how Semantic Web technologies might improve the journalistic practice?
Analysis of the qualitative interviews and the first PD study answers RQ1.

The second PD study examines how the user experience of Danish news journalism can be improved in the context of Semantic Web? **This PD study involves the same group of participants and examines (see Sec.2.6 and Appendix 9.5) potentials of improvement in three news articles. The articles have been randomly selected to represent different genres of news articles (see Table 01 below).**

News article	Genre description	News media	Source
1	Comment response to an ongoing international news event	TV2	https://nyheder.tv2.dk/udland/2020-02-11-who-udbruddet-af-coronavirus-er-meget-alvorlig-trussel
2	Detailed analysis of the same ongoing news event	Berlingske	https://www.berlingske.dk/internationalt/saerligt-en-gruppe-bliver-haardt-ramt-af-coronavirus
3	Reporting on a cultural news event	Danmarks Radio (DR)	https://www.dr.dk/nyheder/kultur/film/historisk-sydkoreansk-film-rydder-bordet-til-oscar-uddelingen

Table 01. News articles selected for PD activities

The scope of this study limits the selection to three articles, even though a broader representation of news article genres would strengthen the research design.
Analysis of the second PD study answers RQ2.

Key findings from the qualitative interviews and the two PD studies are interpreted and examined in the context of Semantic Web. This exploration focuses on technically describing how three specific Semantic Web applications can be developed and implemented to improve the user experience of news articles as well as the work process of news journalists? **The aim of this examination is**

to reveal technical challenges and possibilities and to discuss potentials on a more detailed level. The technical examination answers RQ3.

Finally, Chapter 6.0 of this study returns to the collection of existing Semantic Web applications for the news media industry to compare and evaluate these with key findings of this study.
Analysis and interpretation of the qualitative interviews and the PD studies together with the technical examination and discussion answer the overall problem statement.

2.4 Principles of quality data collection

A common concern about qualitative research is that it provides little basis for scientific generalisation (Kidder & Judd, 1986). To accede to this critic and to establish the quality of any qualitative research study, four tests are commonly used (Kidder & Judd, 1986, p. 26-29). The following paragraphs summaries each test and outline how they are applied in the study at hand.

1. Construct validity: **Establishing correct operational measures for the concept being studied.**

According to Yin (1994), the use of multiple sources of evidence is one strategy to increase construct validity. In this study, different methods of qualitative data collection are applied, and several sources and types of data – such as interview data, and input from PD studies – are collected. In combination with technical analysis, this constitutes multiple sources of evidence as foundation for answering the study's overall problem statement.

Fig.04 Multiple sources of evidence to increase construct validity

2. Internal validity: Establishing a causal relationship, whereby certain conditions are shown to lead to other conditions, as distinguished from spurious relationships (Yin, 1994, p. 34).
This test will not be discussed further as it is only relevant for causal studies.
3. External validity: Establishing the domain to which a study's findings can be generalised.
Qualitative studies rely on analytical generalisation where the investigator is striving to generalise a particular set of results to some broader theory (Yin, 1994, p. 36). This generalisation is however not automatic. A theory must be tested through replications of the findings in a second or even third study, where the theory has specified that the same results should occur. Once such replication has been made, the results might be accepted for generalisation (Yin, 1994, p. 36). In this study, eight participants from four different Danish news media are used as informants (see Sec.2.5 and Table 02), meaning that key findings and recommendation has been tested up against eight participants and four media organisations. This replication logic ensures external validity and makes the study generalisable to Danish news media in a broader sense.
4. Reliability: Demonstrating that the operations of a study can be repeated with the same results.
The objective is to ensure that, if a later researcher conducted the same study all over again, she would arrive at the same findings and conclusions (Yin, 1994, p. 37). To ensure that, thorough documentation of implied methods can be found in Sec.2.5-2.6, and interview guides and descriptions of PD activities can be found as Appendices 9.3–9.6.

2.5 The method of qualitative interview

Interviewing is one of the most frequently used methods in qualitative research and is – when conducted correctly – the most effective way to collect

data about the experiences, opinions, and life worlds of other people (Tanggaard & Brinkmann, 2010, p. 29).

In this study, qualitative interview is chosen to shed light upon the work processes and experiences of Danish news journalists. Idealy, this would be studied through observation, but because of the timeframe – writing news articles can take from a couple of hours to several weeks – it is not possible to collect a substantial and varied data material this way. Instead, qualitative interview is applied to get the participants to share their experiences and habits through expressions, reflections, and narrations.

Interview as a method is not capable of grasping exactly how it is to experience what the participant talks about (Tanggaard & Brinkmann, 2010, p.31), and the qualitative interview researchers Gubrium and Holstein (2003) argues that the method should be considered as a form of interaction between at least two people which results in context based and socially negotiated answers (Gubrium & Holstein, James, 2003). The aim of any qualitative interview, therefore is to get as close to the experiences – in this study the everyday work process of Danish news journalists – as possible and to phrase a coherent and theoretically enriched third person perspective of the experiences (Tanggaard & Brinkmann, 2010, p.31).

The number of interview participants should always be considered in relation to the scope of a research project (Tanggaard & Brinkmann, 2010, p.32). For the scope of this study, two editors and six writing journalists are selected as participants. All of the participants are educated journalists and work for four of largest Danish news media organisations (see Table 02 below), thus they are expected to be conscious of their work process. According to Spradley (1979), good informants should know their culture well and should currently be involved in the cultural scene (Spradley, 1979, p.46), thus a minimum of three years'

work experience has also been a criterion when selecting the participants.

All of the participants work with news journalism, meaning that their work is always based on some kind of factual information – from statistics to information about time and date of a specific event. This makes the participants' work process comparable and allows them to answer the same set of questions (see interview guide in Appendix 9.3). At the same time, the group of participants is assembled to represent different genres of news journalism (see Table 02 below) to ensure that different perspectives and values are included in the study.

Participant	Gender	Years of work experience	News media organisation	Title/work area
1	F	6-10	Berlingske	Reporter, national news and integration
2	F	3-5	Jyllands-Posten	Investigative journalist
3	M	16-20	DR	Data journalist
4	M	11-15	Berlingske	Digital data journalist
5	F	21-25	Berlingske	Digital editor
6	M	3-5	DR	Foreign news desk reporter
7	M	3-5	TV2	Digital editor
8	F	6-10	TV2	Digital culture and lifestyle journalist

Table 02. Participants for qualitative research study

All of the interviews are based on the same interview guide (see Appendix 9.3) and are performed in Danish as this is the working language of all participants.
The interview guide is a translation of a set of research questions targeted at investigating how Semantic Web technologies can support and improve the journalistic work process? (see Appendix 9.3). Each interview question is phrased after Sprad-

ley's (1979) categorisation of descriptive interview questions (see Table 03 below).

Descriptive questions	Intended to encourage an informant to talk about a particular cultural scene
Grand tour questions	Aims at getting a verbal description of significant features of the cultural scene
Mini tour questions	Identical to grand tour questions except they deal with a much smaller unit of experience
Example questions	Take some single act or event identified by the informant and ask for an example, often leading to the most interesting stories
Experience questions	Merely asks informants for any experience they have had in some particular setting

Table 03 Spradley's descriptive questions (Spradley, 1979, p. 83-91)

A grand tour question**:** Can you step-by-step describe how you researched, wrote and published your latest article? **is used as framework for the entire interview and is a translation of the** research question**:** What is the work process of Danish news journalists?

Participants are invited to describe the work process of the news article they have most recently published instead of describing their work in more general terms. The latest article might not be fully representative for the participant's work; however, this technique forces the journalist to answer in more detail and to provide valuable nuanced reflections on specific challenges and opportunities. Answers of the eight participants form a random selection of work descriptions.

To achieve rich descriptions, the grand tour question **is divided into three** mini tour questions **(see Appendix 9.3).** Example questions **are used to encourage participants to be as precise and descriptive as possible, and** experience questions **are employed for the participants to elaborate on those specific examples. The interview guide is followed relatively strictly, but is supported by relevant** follow-up-**,** exploratory**-, and** expository questions **as suggested by Brinkmann and Kvale (2009).**

2.6 Participatory design

Participatory design (PD) developed as a critique of mainstream design and development for not accommodating the multiple voices of future users (Brandt, Binder, & Sanders, 2013, p.146). In the 1980s Scandinavian and North American computer scientists invited workers to mock-up new computer-based tools that extended the skills of workers. Bringing the knowledge of computer systems, scientists, and production workers into productive dialogue with one another called for tools and techniques that could span the gap between knowledge domains (Brandt et al., 2013, p.149).
Today, PD is not one approach but a range of design practices, set of techniques, and tools that support collaborative enquiry into the intertwinement of essential questions such as what to do? and how to do it? when designing or developing new products (Brandt et al., 2013, p.146).

The method is applied in this study because of its ability to engage future users and to bridge technical and domain-specific knowledge. This study contains two small PD studies, one exploring how Semantic Web technologies can support and improve the work process of Danish news journalists? The other one exploring how the user experience of online news articles can be improved through Semantic Web technologies? The two studies consist of selected acknowledged PD techniques which will be described in Sec.2.6.1-2.6.3. The techniques have been selected to support the PD triad of telling, making, and enacting (Brandt et al., 2013).
Both PD studies use the same group of participants as was used for the qualitative interviews (see Table 02) as especially the first PD study (PD I) is a natural continuation of the qualitative interview.
The second PD study (PD II) explores how user experience can be improved, and the participating journalists and editors represent consumers of news journalism. A group of ordinary non-journalist consumers could also have

been used; however, the journalists and editors add valuable editorial insights to the study on top of their ordinary needs and wishes as news consumers.

Often, PD studies are carried out as focus group interviews with a number of participants interacting in activities at the same time. Location of the participants and the scope of this study with only one reserach do however not provide for at focus group set-up. Instead one participant at time is asked to engage in the same set of PD activities (see Sec.2.6.1–2.6.3) allowing the researcher to collect empirical data for later analysis.

A preceding test of the qualitative interview and PD studies have been carried out. A valuable finding was to undertake the PD studies in relative continuation of the qualitative interview as participants can then draw on challenges and possibilities described in the interview when engaging in activities in the PD studies.

The following sections provide detailed descriptions of the PD activities applied. To see how the activities are combined and presented, please see Appendices 9.4–9.5.

2.6.1 The future workshop

The future workshop is a robust and relatively simple technique which aims at expanding the dialogue between designers (researchers) and users (Brandt et al., 2013, p. 152). The method is chosen for this study because of its ability to introduce a change perspective to shed new light on the well-known.

Participants are invited to list points of critique to their present situation, before they are given the opportunity to develop a utopian perspective. In the

last phase, the utopian vision forms base for a plan of action, where participants discuss what can be done to move towards the vision (Brandt et al., 2013, p. 152). As an example, participants in this study are asked to formulate specific search queries relevant to their latest work which cannot be answered by standard search engines (see Appendix 9.15).

The technique is applied as framework for PD I (see Appendix 9.4) and draws on information from the qualitative interview, where participants were invited to describe obstacles and challenges (points of critique) in their current work process. Before engaging in the activity, participants receive a short introduction to the concept of Semantic Web including specific use cases (see Appendix 9.6). The introduction is designed to give participants a broad understanding of the concept but without affecting their answers and reflections more than absolute necessary, e.g. use cases of journalistic products or applications are not included in the presentation.
The aim of the activity is to gain insights in how journalists perceive Semantic Web technologies, what potential use cases they envision, and what challenges they connect to these.

2.6.2 Scenarios

The scenario technique has long been recognised as powerful and has been widely used in PD practices within HCI (Brandt et al., 2013, p.166). According to Bødker et al. (2004), the technique supports the building of coherent visions and the anchoring of these:

> Scenarios visualise the practical application of a proposed IT system, that is, the potential effects of implementing it. Scenarios are prose-style representations exemplifying a work practice under future use of the system. (Bødker, Kensing, & Simonsen, 2004, p.261)

The second activity in PD I invites participants to evaluate (telling) and engage (enacting) in current and thought-up examples of Semantic Web applications. The selected examples can be seen as scenarios providing specific and visual descriptions of how Semantic Web technologies can be applied in the media industry.

Two of the examples are very early-stage examples of how Semantic Web technologies can enrich information. These are chosen to test whether the journalists are able to recognise the value and perspective in combining multiple datasets. The third use-case is a thought-up example and is easier for the participants to translate into their own domain.

2.6.3 The Magic If and Prototypes

The last activity in PD I draws on the concept of the magic if which was introduced by the actor-director-teacher Stanislavskij who believed that the little world if is what initiates all kinds of creative processes. Brandt and Grunnet (2000) argues that PD practices too can benefit from the world of drama as it is important to explore the context of use from new perspectives (Brandt & Grunnet, 2000, p.11).
Participants are asked to imagine a work setting where their computer is able to perform anything imaginable (see Appendix 9.4). The activity is an experiment of thought and is included to facilitate the broadest possible level of imagination among the participants.

PD II invites to a higher level of making and enacting by including simple prototypes and elements of drama. According to Marion and Suri (2000), experience prototypes are any kind of representation, in any medium, that is designed to understand, explore, or communicate what it might be like to engage with the product or system that is being designed (Buchenau & Suri, 2000, p.426).

In this study, prototypes are being created throughout PD II (see Appendix 9.20), as participants are invited to draw on top of existing online articles and make simple prototypes to improve the user experience (see Appendix 9.4). First, participants are restricted to apply improvements within the framework of Semantic Web, thereafter they are free to add any type of addition.

2.7 Coding transcripts for analysis

Transcripts of qualitative interviews and PD studies have been coded and categorised (see Appendices 9.7–9.14) inductively after a grounded theory approach, meaning that instead of using prescribed themes or theories to define categories and subcategories, these have been identified throughout the coding process. This approach is chosen because this part of the analysis is not based on a firm theory or philosophical view. Instead the empirical material is expected to reveal patterns and relations, thus it makes sense to let the data dictate the number and names of categories.

In total, eight categories are identified across the transcripts (see Table 04 below). The coding process has been conducted on transcripts of one participant at the time – coding first the qualitative interview, then the two PD studies – before starting all over with the next set of transcripts. This way, coherence between the qualitative interview and PD studies is kept during coding and analysis. The fact that all coding categories are identified in both the qualitative interviews, PD I and PD II is also an argument for coding and analysing all transcripts as one dataset.

Category	Category description
Indexing and archiving	References describing the importance of indexing and archiving published articles and the journalist's own research

common archive	References describing how a media uses tags and other techniques to index and archive published articles and research
personal archive	References describing how journalists archive research and published articles
Overview	References describing the importance of creating overview of information when researching as well as when presenting knowledge to the users
research	References describing how journalists search in previously published news articles and other sources as part of the research process
for users	References describing how journalistic services is used to create overview for users
Deadlines and efficiency	References describing time pressure and the importance of efficiency in the journalistic process
Trust and reliability	References describing the importance of trust and reliability during research and when publishing news articles
towards sources	References describing the importance of reliability and trustworthiness of external sources in the journalistic process
towards the media	References describing how a media's trustworthiness can be maintained and enforced
Query examples	References describing participants' proposals of relevant Semantic Web queries
Sources and contacts	References describing how journalists find and use sources and their contact details
Graphical presentation	References describing the importance of graphical presentation of data and how it can be improved
Interactive and user-oriented services	References describing the importance of interactive and user-oriented services and how it can be improved
Social media search	References describing the need and enquiries for searching in social media resources

Table 04 Listing of coding categories and category descriptions

Further analysis and results of the qualitative research studies can be found in Chapter 4.0.

3.0 Theoretical framework

Chapter 3 presents relevant theories for answering the problem statement of this study. Sec.3.1 summarises the concept of Semantic Web and explains each layer of the Semantic Web Stack. In Sec.3.2 principles of linked data are presented, before Sec.3.3 in more detail describes concepts of Resource Description Framework (RDF) including serialisation. Sec.3.4 elaborates on the use of web ontologies, while Sec.3.5 touches upon the methodology of querying linked data. Finally, Sec.3.6 outlines important aspects of Unifying Logic, Proof, and Trust and the state of research for these top layers in the Semantic Web Stack.

3.1 Concept of Semantic Web and the Semantic Web Stack

In broad terms, the Web as we know it builds on the following three principles:

- A worldwide addressing schema enabling each web document to have a unique identifier – an URL
- A transport layer – the HTTP – which supports remote access to content over a network layer – TCP-IP
- And a platform-independent interface – HTML and Web browsers – which enables users to easily access any online resource
  (Domingue et al., 2011, p. 8)

According to Berners-Lee et al. (2001), Semantic Web builds upon these principles and can best be understood as:

> An extension which gives information well-defined meaning and better enables computers and people to work in cooperation.
> (Berners-Lee et al., 2001).

This extension aims at tackling problems of accessing data and enabling delegation which are current issues central to the Web (Domingue, Fensel, & Hendler, 2011, p.9).

Complex matching or querying is not possible with current search engines which illustrates the first problem of accessing data: If a journalist wants to know what members of the Danish parliament have studied law, the journalist needs to look up each member[4], find information about their educational background, and combine this information. Even though the information is accessible, standard search engines and the Web in general are not capable of making this integration of information.

The issues of enabling delegation is illustrated every time journalists (or other users) browse the Web. Then computers act simply as rendering devices displaying text, audio, and other content, whereas all inferences and computation are left for humans to do (Domingue et al., 2011, p.10). Inference and integration of information is usually very timeconsuming, and it often hampers the work process or even hinders journalists in doing specific types of research and investigation (see Sec.3.4).

Currently the Web is loaded with accessible information, but great potentials of easily combining different datasets or even autogenerating new information based on accesible data can not be fulfilled as data today is not machine-readable. To achieve those potentials, Semantic Web is crucial:

> Delegating tasks such as the integration of information, data analysis, and sensemaking to machines, at least partially, is the only way forward for users, communities, and businesses to continue to make the most of the information available on the Web.

4 Each member of the Danish parliament has a profile page describing their professional lives, contact details, and political achievements, e.g. https://www.ft.dk/medlemmer/mf/n/nick-haekkerup.

(Domingue et al., 2011, p.10).

To attain machine-readability, Semantic Web reuses the Web's global indexing and naming scheme, meaning that in principle every semantic concept has a unique identifier (Domingue et al., 2011, p.5). On top of this, semantic technology adds annotations to semi structured information as database technology adds column headings to tabular information. This can be illustrated as in the syntax below:

```
<person>
   <name>Nick Hækkerup</name>
   <profession>member of parliament</profession>
   <education>law</education>
</person>
```

The annotation above allows a computer to understand that Nick Hækkerup is a name of a Person, and that this person has a Profession as member of parliament and an Educational background in law. Based on this annotation – and if the description of all other members of parliament where annotated similarly – the search question in the example mentioned above can easily be queried and answered (see Fig.05 below). This type of search is referred to as Semantic Web query.

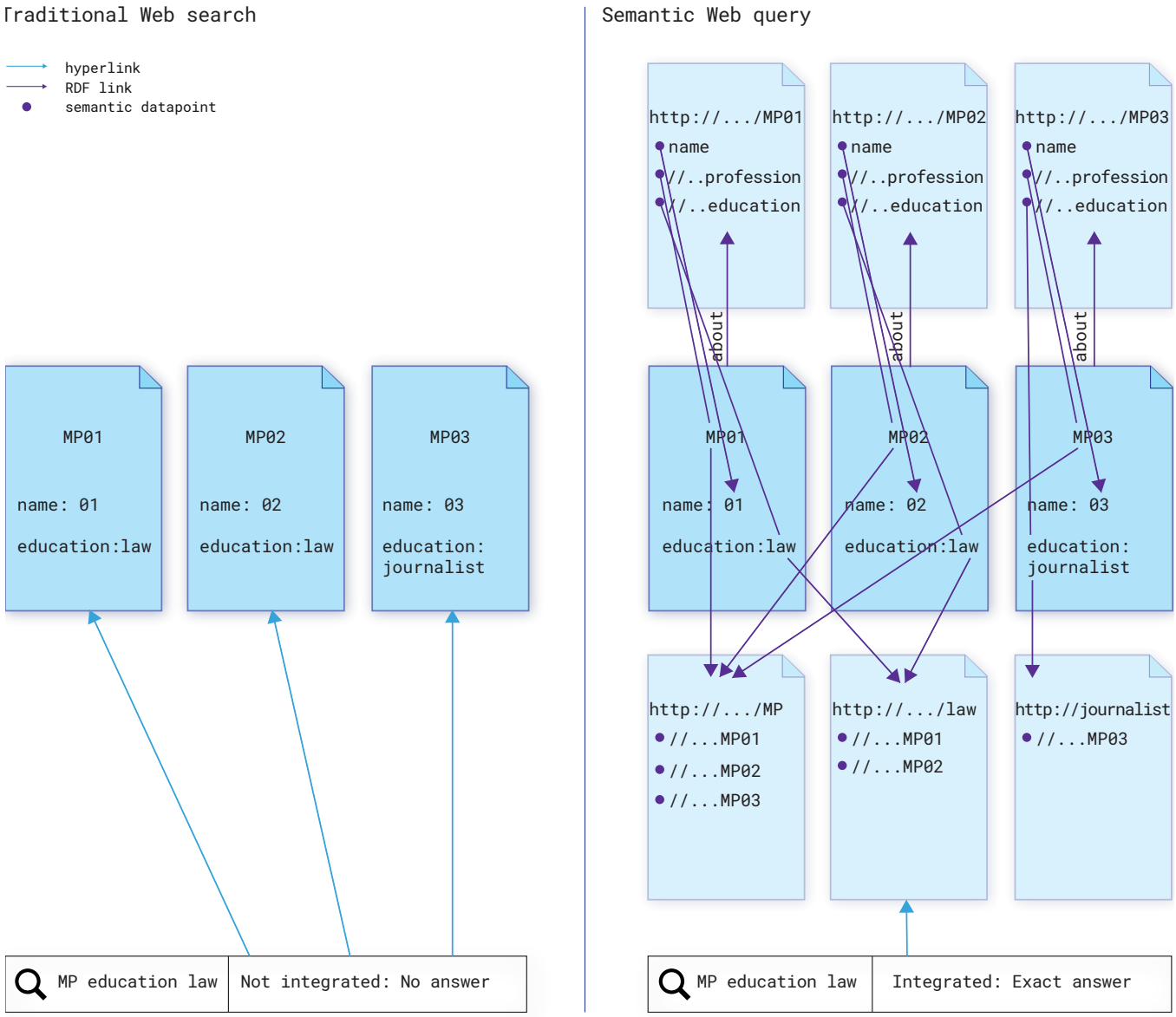


Fig.05 Showing traditional search on the Web (left side) versus search on the Semantic Web or a Semantic Web query (right side). When information is semantically annotated, it is possible to get an exact answer to the query: Which members of parliament has studied law? This is not possible on the traditional Web, as machines do not understand the semantic meaning of traditional web documents.

In very broad terms, two things are needed to define the semantics of information: A language – such as <x>Y</x> in the syntax example above – to define the meaning, and terms – such as x in the syntax example above – to denote this meaning (Domingue et al., 2011, p.13).

In the concept of Semantic Web, the terms – or meaning – is considered a shared resource:

> (...) Semantic Web incorporates the notion of an ontology, which by definition is a shared machine-readable representation.
> **(Domingue et al., 2011, p.5).**

Through ontologies and ontology-related technologies (see Sec.3.5), the meaning and relationships between concepts published on the Semantic Web can be processed and understood by software-based reasoners (see Sec.3.6). These principles are illustrated in the Semantic Web Stack model (Fig.06 below).

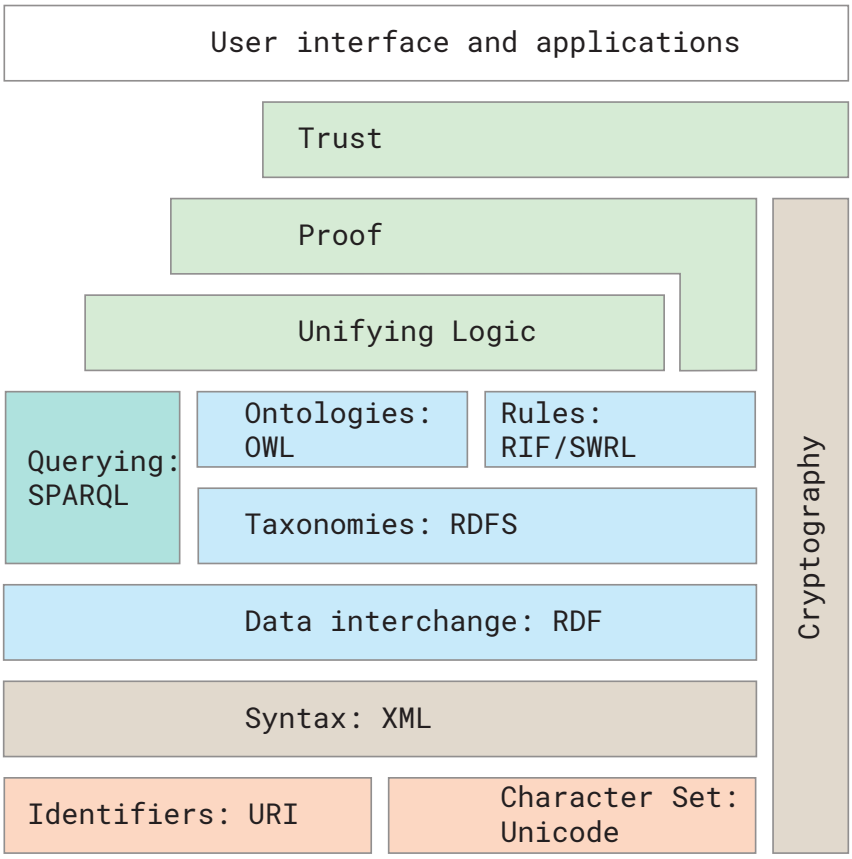


Fig.06 The Semantic Web Stack (after Berners-Lee, 2006) illustrates how one technology builds on another to constitute the Semantic Web.

The bottom layers up until the layers of Unifying Logic, Proof, and Trust are formats standardised by the W3C Consortium (Koivunen & Miller, 2001) while the top layers to some extend are still matters of research and contain fundamental challenges of how to ensure documentation and transparency for

statements generated by the Semantic Web.
The remain of this section outlines the groups of layers, while Sec.3.2–3.6 describe the principles and technologies contained in each layer in more detail.

The bottom layers – the Uniform Resource Identifier (URI) **and the** Unicode[5] **layers – constitute the foundation of Semantic Web and are already implemented when writing web documents with user-defined vocabularies (Alam, Rahman, Khusro, & Ali, 2015, p.116).**
The Extensible Markup Language (XML) **layer allows Semantic Web definitions to be integrated with other XML based standards and provides means for uniquely identifying concepts in Semantic Web (Koivunen & Miller, 2001).**
In the Resource Description Framework (RDF) **and** Resource Description Framework Shema (RDFS) **layers, resources and links are provided with types. RDF makes it possible to create URI-defined statements about each concept, and RDFS includes vocabularies that can be referred to also by URI's (Koivunen & Miller, 2001).**
The Ontologies **and** Rules **layers support the evolution of vocabularies which define relations between different concepts.**
The Cryptography **layer (to the far right in Fig.05 above) detects alterations to documents, and the** Querying **layer (to the far left in Fig.05) is where linked data are fetched for possible reuse (Koivunen & Miller, 2001). This can be done using** SPARQL Protocol And RDF Query Language (SPARQL)**.**
The top layers consist of technologies – including Unifying Logic**,** Proof **and** Trust **– which are not yet standardised. The aim of these layers is to provide declarative knowledge and proof of validation to gain users' trust for its operations and the information provided.**
Finally, users interact with Semantic Web application through User interfaces

5 Unicode is a universal character encoding standard designed to define the way characters from all human languages are represented in documents. There are several types of Unicode encodings, though UTF-8 is the most common (Techterms, 2019).

built on top of the Semantic Web Stack (Alam et al., 2015, p.116). Sec.3.5 which briefly mentions one generic method of implementation, while application specific implementation is further discussed in Chapter 5.0.

3.2 Principles of Linked Data

In order for the Web of linked documents to evolve into a Web of linked data, Berners-Lee introduced a set of best practices – the Linked Data Principles – for publishing and connecting structured data on the Web:

1. Use URIs as names for things
2. Use HyperText Transfer Protocol (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 and digital content, but also real-world objects and abstract concepts. URIs can be used to identify even small objects such as locations or mobile numbers and describe their metadata relationships. Unicode in combination with URI extends support for identifying any type of resource regardless of its text and scripting language (Alam et al., 2015, p.116).

The second Linked Data Principle advocates the use of HTTP URIs to identify objects and abstract concepts just like the HTTP protocol is the universal access mechanism for the traditional Web (Berners-Lee, 2006). According to Hendler, Heath & Bizer (2011), HTTP URIs make good id's for two reasons:

— They serve not just as a name but also as a means of accessing information

describing the resource,

— and every owner of the domain name may create new URI references

(Hendler, Heath, & Bizer, 2011, p.10).

The third principle advocates that HTTP clients should be able to look up any HTTP URI and retrieve a description of the resource. In order to make different statements about an object and about the document describing that object, it is common practice to use different URIs to identify the real-world object and the document that describes it (Hendler et al., 2011, p.10). Thus, each entity will likely have at least three URIs (see Table 05 below).

Types of URI-reference	Example
URI for the real-world object	http://dbpedia.org/resource/news_article
URI for a related information resource that describes the real-world person and has an HTML representation	http://dbpedia.org/page/news_article
URI for a related information resource that describes the real-world person and has an RDF/XML representation	http://dbpedia.org/data/news_article

Table 05 Types of URI-references

URI's used for declaring vocabularies are called namespaces. These are often associated to prefixes instead of the full URI (Domingue et al., 2011, p.124).

The agreement of HTML as the dominant document format has been crucial for Web's ability to scale (Hendler et al., 2011, p.11). Similarly, it is important to agree on a standardised content format for URI descriptions. According to the third Linked Data Principle, this format should be RDF which is readable for both humans and machines (Berners-Lee, 2006).

Each URI description should include the following RDF triples (see Sec.3.3):

- **triples that describe the resource with literals**
- **triples that describe the resource by linking to other resources**

- triples that describe the resource by linking from other resources
- triples describing related resources
- triples describing the description itself
- triples about the broader data set of which this description is a part (Hendler et al., 2011, p.45)

The fourth principle advocates the use of hyperlinks to connect not only Web documents, but any concept described on the Web (Berners-Lee, 2006). This way it is possible to interlink a political proposal to the politician who wrote it. Furthermore, that politician can be linked to his or her political party, colleagues and so on. Such hyperlinks in a linked data context are called RDF links.

In summary, the Linked Data Principles lay the foundation for extending the Web with a global data space based on the same architectural principles as the traditional Web. The following sections explain the technical realisation of the Linked Data Principles in more detail.

3.3 Resource Description Framework (RDF)

The Resource Description Framework (RDF) is a data model for publishing statements on the Web. The data model is designed for an integrated representation of information that

— originates from multiple sources
— is heterogeneously structured
— and is represented using different schemata (Hendler et al., 2011)

In RDF, a description of a resource is represented as a number of triples[6]. Each triple consists of a subject, a predicate, and an object which mirrors the struc-

6 Such triples are also referred to as statements.

ture of a simple statement:

Nick Hækkerup	**works**	**as a politician**
subject	predicate	object

The subject of a triple is the URI identifying the described resource. The object can either be a simple literal value[7] 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. Predicate URIs come from vocabularies or ontologies (see Sec.3.4). There are situations, when only the type of object is known, or when the instance is anonymous. The RDF response to this is so-called blank nodes (Hendler et al., 2011, p.17).

As described in Table 06 below, two principal types of RDF triples can be distinguished:

Type of RDF triple	Object-subject-predicate relation	Function
Literal triple	has an RDF literal as the object	is used to describe properties of resources (such as the nickname or dateOfBirth of a Person)
RDF link	consists of three URI-references as object, subject, and predicate	describes the relationship between two resources
Internal RDF link	subject, object, and predicate consist of URI-references within the same namespace	connects resources within a single linked data source
External RDF link	subject of the triple is an URI-reference in the namespace of one data set, while the predicate and/or object are URI-references pointing into the namespace of other data sets	connects resources served by different linked data sources

Table 06 Types of RDF triples after Hendler et al., 2011, p. 16-20.

7 A literal value can be a text string, a number, a Boolean, or a date.

External RDF links are crucial for the Semantic Web as they glue different data islands into a global, interconnected data space (Hendler et al., 2011, p.16). This concept is supported by the graph representation of RDF (see Fig.07 below as a segment example). Semantic Web applications operate on top of this giant global graph and retrieve parts of information by dereferencing URIs required.

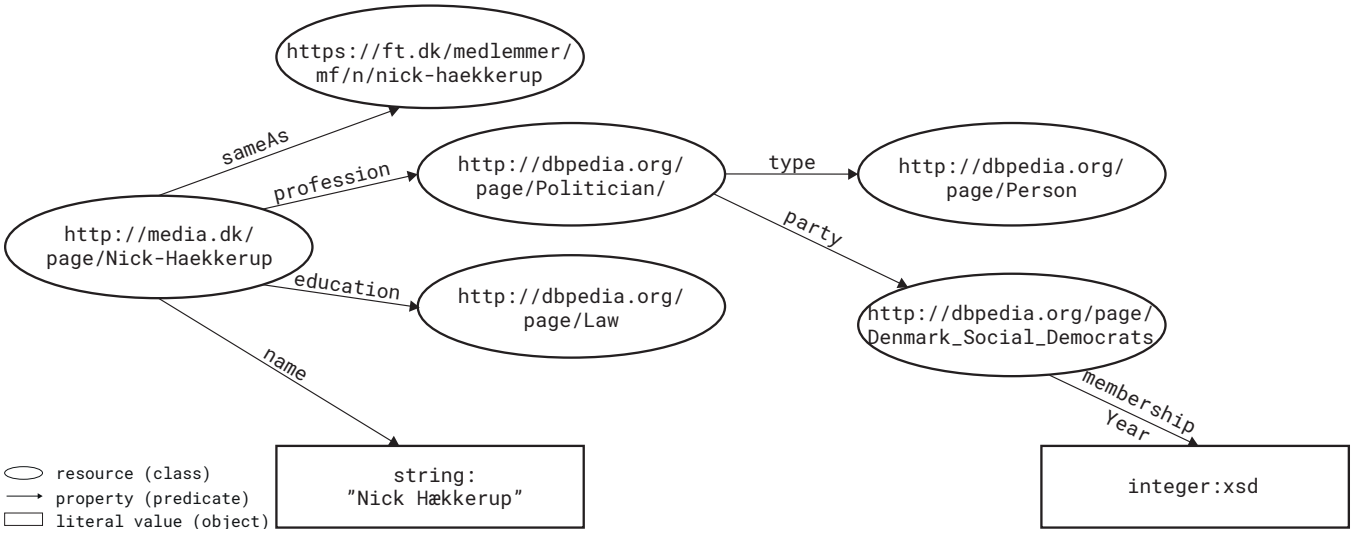


Fig.07 RDF graph model combining information from three different online databases. It is possible to imagine all linked data as one giant global graph.

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, p.146).

Another important aspect of the logical view of RDF triples is that RDF makes an open-world assumption, meaning that RDF semantics assume that whatever is not explicitly stated could be true, while in relational models the assumption is, that facts that are not explicitly claimed are false (Domingue et al., 2011, p.122).

3.3.1 RDF Serialisation

As RDF is a data model and not a format, an RDF graph must be serialised to be published on the Web. This section describes the two formats that have been standardised to do so by the W3C: RDF/XML and Resource Description Framework in Attributes (RDFa).

The aim of RDF/XML is to be machine-processable and compliant to XML which allows RDF documents to be easily exchanged between different types of applications (Domingue et al., 2011, p.126). The listing below shows the RDF/XML serialisation of two RDF triples.

```
1  <?xml version="1.0" encoding="UTF-8"?>
2  <rdf:RDF
3      xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
4      xmlns:shcema="http://schema.org/">
5
6      <rdf:Description rdf:about="http://media.dk/resource/person/nick-haekkerup01">
7          <rdf:type rdf:resource="http://schema.org/Person"/>
8          <schema:name>Nick Hækkeup</schema:name>
9      </rdf:Description>
10
11 </rdf:RDF>
```

Snippet 01 RDF/XML serialisation of two RDF triples. The first triple states that there is a resource, identified by the URI http://media.dk/resource/person/nick-haekkerup01 of the type Person. The second triple states that this Person has the name Nick Hækkerup.

RDF/XML is hard for humans to read and write, and workflows that involve human intervention might be problematic. RDF/XML can also be very verbose, hindering expressivity (Domingue et al., 2011, p.126).

RDFa is a serialisation format that embeds RDF triples in the HTML document (see Snippet 02 below), meaning that existing content within the HTML code can be annotated with RDFa (Hendler et al., 2011, p.19). When using RDFa, it is important to maintain the unambiguous distinction between real-world objects and the HTML-RDFa document that embodies descriptions of these. This

can be achieved by using the RDFa attribute about= to assign the relation between a Web document and the real-world objects that are being described within that document (see Fig.17).

```
<!DOCTYPE html>
<html   rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
        schema="http://schema.org/">

<head>
    <meta http-equiv="Content-Type" content="application/xhtml+xml;
        charset=UTF-8"/>
    <title>News article about Nick Hækkerup</title>
</head>

<body>
    <div about="http://media.dk/resource/person/nick-haekkerup01" typeof="schema:Person">
        <p property="schema:name">Nick Haekkerup</p>
    </div>
</body>

</html>
```

Snippet 02 RDFa serialisation of the same two RDF triples as described in Snippet 01.

A number of individuals and organisations have already adopted the Linked Data Principles (see Sec.3.2). One example is DBpedia[8], which is a dataset extracted from publicly available Wikipedia dumps. Because of its breadth of topical coverage, DBpedia has served as a hub within the Semantic Web since 2007 (Hendler et al., 2011, p.33).

As described in Sec.1.1.1, several services have also been launched to automatically or semi-automatically annotate text documents including Calais and N.Y.T.'s Editor. The practical implementation of semantic annotation is however still considered one of the greatest challenges to the realisation of Semantic Web (Domingue et al., 2011).
Implementation of semantic annotation is discussed further in Sec.5.2.2.

8 http://www.dbpedia.org

3.4 Ontologies

RDF provides a data model for describing resources, but it does not provide any domain-specific terms for describing classes[9] **of things, and how they relate to each other. This function is served by lightweight ontologies which are particularly useful for representing knowledge from unstructured text because of their flexibility and ability to evolve:**

> (...) once created, ontologies can be far more easily extended than is for example the case for relational database schemas.
> **(Domingue et al., 2011, p.755)**

Most often ontologies in the context of Semantic Web are expressed in RDFS (Hendler et al., 2011, p.57) to declare hierarchies and relationships. As an example, an ontology can define that the class Politician **is a subclass of class** Person**, which is expressed in Description Logics[10] (DL) below:**

```
Politician ⊑ Person
```

rdfs:subClassOf **is used to state that all instances of one class are also instances of another, meaning** that entities of the subclass inherits all properties and taxonomies of the dominant class **(Hendler et al., 2011, p.57).**

Properties are used to state the relation between two resources or between a resource and a literal value: The resources Nick Hækkerup **and** Law **can be related by the property** hasStudied **(stating that Nick Hækkerup has studied law):**

```
(Nick Hækkerup,Law) hasStudied
```

Similarly to classes, properties can also be hierarchically described by rdfs:subPropertyOf**, which states that all resources related by one property are**

9 The word class is used in ontologies instead of term or category to describe a category of entities.
10 Description Logics are a family of formal knowledge representation languages often used in AI-research to describe and reason about the relevant concepts of an application domain.

also related by another.

rdfs:domain is used to state that any resource that has a given property is an instance of one or more classes, and rdfs:range is used to state that all values of a property are instances of one or more classes (Hendler et al., 2011, p.57). E.g. the class Person is domain for the relation hasStudied, and that the class Education is range for that relation:

$\ni$hasStudied.T$\sqsubseteq$Person
T$\sqsubseteq\forall$ hasStudied.Education

Classes and properties are themselves resources, whose URIs according to the Linked Data Principles should be made dereferenceable (Hendler et al., 2011, p.57). Existing standard vocabularies (see Sec.5.2.1) should be applied wherever possible as underlying ontology, taxonomy and rules can then be reused (Hendler et al., 2011, p.61).

Authors of RDFS vocabularies define rules that when paired with a suitable reasoning engine (see Sec.3.6) enable implicit relationships and additional information to be inferred (Hendler et al., 2011, p.57). Uschold and Gruninger (2004) mentions neutral authoring as a way to achieve common understanding through ontologies. Interestingly they do not describe what the adjective neutral means. Instead, they go on to add that:

> (...) the neutral ontology must cover all of the concepts in each of the target systems.
> **(Uschold & Gruninger, 2004, p.62)**

This description can be seen as what others refer to as an overarching ontology[11] (Hendler et al., 2011), but it is still unclear what neutrality is and how it is achieved. This is further discussed in Sec.6.3.

11 An overarching ontology subsume multiple database schemas based on open, lightweight standards (Hendler et al., 2011, p. 749).

3.5 Querying data

The Linked Data Principles allow for publishing and accessing simple facts but do not support more complex queries (Domingue et al., 2011, p.56). To retrieve this type of information, SPARQL must 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, chap.5).

A SPARQL-query consists of clauses that define different aspects of the query. Table 07 below describes some of the most common used clauses.

Clause	Function
PREFIX	abbreviates URIs for clarity and to improve readability of the graph pattern
FROM	declares what RDF graph(s) is/are being queried
SELECT	specifies information interest
CONSTRUCT	is used to produce new RDF graphs that either can be presented to the user or fed as data source to the next query
WHERE	defines the exact graph pattern that has to be matched against Semantic Web data. As described in Sec. 3.3 a basic graph pattern consists of individual (subject, predicate, object) patterns which are joined by variables, forming a template that will be filled during the matching process
FILTER	narrows returned results to structures that fulfil specific criteria

Table 07 Common SPARQL clauses and functions after Domingue et al., 2011

Each clause and the combination of these must follow a specific syntax which is illustrated in Snippet 03 below.

```
PREFIX dbo:<http://dbpedia.org/ontology/>
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
PREFIX dbp:<http://dbpedia.org/property/>
PREFIX dct:<http://purl.org/dc/terms/>
PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl:<http://www.w3.org/2002/07/owl#>

SELECT ?name ?description
WHERE{
    ?politician a dbo:Politician.
    ?politician dbo:hasEducation dbr:Law.
    ?politician rdfs:label ?name.
    ?politician dbo:abstract ?description.
FILTER (lang(?name)="en").
    }
```

Snippet 03 SPARQL-query requesting name and description of politicians who have been studying law. In this example, information is received from the database of open linked data DBpedia

Datasets can be exposed via a SPARQL endpoint[12] accessible via HTTP requests which enables remote access to the data and frees it from closed data silos (Domingue et al., 2011, p.60).

SPARQL results can be implemented differently in a wide range of applications. One way is to format the response in JavaScript Object Notation (JSON) and use the REpresentational State Transfer (REST) request model which is one of the most used web service technologies allowing two computer systems to communicate over HTTP. The model decouples data providers from data consumers and allows clients to use uniform interfaces for accessing and querying data stored on remote servers (Domingue et al., 2011, p.55).

In the RESTful model, information about resources identified by URIs can be accessed directly using simple HTTP GET method[13] (see Snippet 08).

12 An example of a SPARQL endpoint can be found here: http://sparql.org/sparql.html

13 GET is used to request data from a specified resource and is one of the most common HTTP methods. Data is encoded into the URL and appended to the action URL as query string parameters (W3C, ref_httpmethods).

3.6 Research on Logic, Proof, and Trust

This section returns to the top layers of the Semantic Web Stack: The Unifying Logic, Proof and Trust layers which are crucial to most Semantic Web applications as it is fundamental to know where presented information comes from and how resulting conclusions have been constructed. This is not an issue to the current Web as humans do most data analysis and integration. Issues of documentation and trustworthiness will however be amplified as descriptive reasoners and other Semantic Web agents play a greater role (Jarhi & Gaaly, 2007). Thus, mechanisms for automatic proof checking must be included. The layers contain technologies that are not yet standardised, and research within each area lacks to provide answers to several questions.

The Unifying Logic layer consists of automatic reasoning systems which operate on top of the ontology structure to make new inferences (Pandey & Sanjay, 2010, p. 30). It is out of the scope of this study to describe the different types of logic that perform these inferences, but it is worth briefly mentioning the Web Ontology Language (OWL) and its sublanguages OWL Lite – which provides first order logic – and OWL LD – which provides descriptive logic. These sublanguages have the capability to deduce complex knowledge from ontologies (Alam et al., 2015, p. 120). E.g. if a thing of the type Person has the property profession="memberOfParliament". Then that Person must also be a Politician which is a subclass of Person and is the domain of the property memberOfParliament.

This way, Semantic Web technologies not only provide access to combined and integrated information, they also enable more decentralised knowledge-based information to be constructed which ultimately empowers more advanced systems such as Artificial Intelligence (AI) (see Sec.6.3).

Unifying Logic is currently implemented in several Semantic Web applications,

but standards to ensure transparency with how applied ontologies and reasoning mechanisms are constructed are still missing and require more research. Without such transparency, biased or manipulated ontologies can ultimately provide answers which cannot be distinguished as true or false.

Berners-Lee (2006) included the Proof layer to describe for software agents (or human users) why they should believe a retrieved result (Pandey & Sanjay, 2010, p.30). Generally, metadata about a resource is considered as a proof of its content.
Sergej (2007) defines this as different types of provenance which originally is an approach applied in research on relational databases. Provenance describes how information arrives as a query answer, including information about RDF statements contributing to produce the answer and how the query result was produced (Sizov, 2007, p.95). Provenance must be considered for each piece of metadata and for each triple.

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, and license (Hendler et al., 2011, p.48). One mechanism for publishing this type of metadata is Semantic Sitemaps which are an extension of the well-established Sitemaps protocol.
Sitemaps protocol hints about Web pages that are available for crawling and consists of an XML document stored in the root directory which defines elements such as URL, loc, lastmod, and changefreq (Domingue et al., 2011, p.48). This enables data publishers to inform semantic software agents where they can retrieve additional descriptive information in RDF (Hendler et al., 2011, p.48).
In practice however, semantic metadata are processed by different loosely coupled systems which makes tracking, propagating, and querying difficult (Jaques et al., 2012, p.37).

The Trust layer has not yet progressed far beyond a vision of allowing people to ask questions about the trustworthiness of information on Semantic Web (Pandey & Sanjay, 2010, p.30):

> The trust layer shall ensure that the source of information is judged, and also this layer shall ensure that only authorised applications/agents and authorised users only have an access to the information.
> **(Pandey & Sanjay, 2010, p.30)**

As suggested by Jarhi & Gaaly (2007), a Web of Trust can be attained through digital signatures which are envisioned to check if data really comes from the claimed and trusted source. Digital signatures can also be used in combination with encryption to ensure confidentiality of information (Jarhi & Gaaly, 2007). Implementation of such technologies and systems requires further research.

As discussed above, it is crucial to ensure reliability and trustworthiness within Semantic Web applications. Current research envisions how this 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 Semantic Web is fully trustworthy as the layers of Trust and Proof have not yet been developed to an acceptable extend.
Aspects of trustworthiness and reliability is further discussed in Sec.5.2.4, 5.3.4, 5.4.4, and 6.3.

4.0 Analysis: Creating news journalism

Chapter 4 provides an overview of key findings from the qualitative analysis carried out on participant interviews and PD studies.
Sec.4.1 briefly outlines how the following sections are organised in relation to the analysis. Sec.4.2–4.5 present key findings and interpretations as well as suggestions for valuable Semantic Web applications. Detailed documentation for these conclusions is also provided. The chapter concludes with a partial conclusion in Sec.4.6 answering RQ1 and RQ2.

4.1 Inductive analysis

As mentioned in Sec.2.7, transcripts of qualitative interviews and PD studies of eight participants have been inductively coded (see Appendices 9.7–9.14) and thereafter analysed.
In total, 16 coding categories have been identified – these are listed with category name and description in Table 08 below.
The table's third column shows the total number of references (quotes from the participant transcripts) identified for each category. Indexing and archiving is the category with the most references – a total of 76 – whereas only 11 references constitute the category Social media search.
The last column shows in how many sets of transcripts each category is identified – 8/8 meaning that the category has been identified across all eight sets of transcripts of participant interviews and PD studies. This documents not only whether a category is represented by a high number of references, but also whether these references come from only a few participants or if the topic is mentioned by all of the journalists and editors.

Category	Category description	Total number of references	Files in which the category is identified
Indexing and archiving	References describing the importance of indexing and archiving published articles and the journalist's own research	76	8/8
the media's archive	References describing how a media uses tags and other techniques to index and archive published articles and research	40	7/8
personal archive	References describing how journalists archive research and published articles	17	6/8
Overview	References describing the importance of creating overview of information when researching as well as when presenting knowledge to the users	76	8/8
research (in other sources)	References describing how journalists search in previously published news articles and other sources as part of the research process	44	7/8
for users	References describing how journalistic services is used to create overview for users	32	8/8
Deadlines and efficiency	References describing time pressure and the importance of efficiency in the journalistic process	51	8/8
Trust and reliability	References describing the importance of trust and reliability during research and when publishing news articles	51	8/8
towards sources	References describing the importance of reliability and trustworthiness of external sources in the journalistic process	38	8/8
towards the media	References describing how a media's trustworthiness can be maintained and enforced	10	7/8
Query examples	References describing participants' proposals of relevant Semantic Web queries	47	8/8
Sources and contacts	References describing how journalists find and use sources and their contact details	41	6/8
Graphical presentation	References describing the importance of graphical presentation of data and how it can be improved	24	6/8
Interactive and user-oriented services	References describing the importance of interactive and user-oriented services and how it can be improved	15	4/8
Social media search	References describing the need and enquiries for searching in social media resources	11	4/8

Table 08. Coding categories – references

The category Query examples is analysed separately (see Appendix 9.15). This category consists of specific examples of Semantic Web queries (see Fig.05 in Sec.3.1 for the difference between traditional Web search and Semantic Web queries) proposed by the eight participants throughout the interviews and PD studies. These are expressions of what the participants themselves wish they could use semantic search or similar technologies for, e.g.:

> Which municipalities are led by a mayor representing Venstre (liberal party in the Danish Parliament, red.)? Or even... if I were to be truly clever, I would probably ask: Danish mayors presented by their political party. **(Participant 04)**

A total of 22 examples have been collected and re-coded after the existing categories. 12 query examples concern the process of finding information in other data sources and especially in previously published news articles. Seven query examples concern the process of organising published information in a suitable way to provide overview for users. Two examples concern challenges of finding contact details, and one query example concerns the creation of graphic material.
18 out of the 22 hypothetical query examples rely on open data sources such as geographical data, information from the media's own articles, or statistic information from Danmarks Statistik[14].

As a point of departure, coding categories, query examples, and common references from the empirical material have been organised in flow diagrams (see Appendix 9.16–9.19).
The diagrams reveal four areas of interest which are presented in more detail in the following sections. The scope of this study does not allow all aspects of the analysis to be unfolded. Themes such as graphical presentation, social

14 Danmarks Statistik or Statistics Denmark is the central authority on Danish statistics.

media search, and user-oriented services are discussed throughout the interviews and PD studies but have been left out of the analysis as the potential for Semantic Web applications within these fields seem less obvious.
All references (quotes) are translated to English in the following sections. To see the original quoting in Danish please refer to the right set of transcripts in Appendices 9.7–9.14.

4.2 Search for contact details

51 references in the transcript material concern time pressure and the need of an efficient work process. In the process of creating online news, journalists work with constant deadlines and in competition with other media on larger news stories. Participants describe how they sometimes spend only a few hours producing a news article, and that deadlines in combination with time-consuming activities are what limit their work. Thus, every tool or technique that can make the work process more time efficient is of value. On the other hand, new services or applications must not require more time for the journalists to maintain or use.

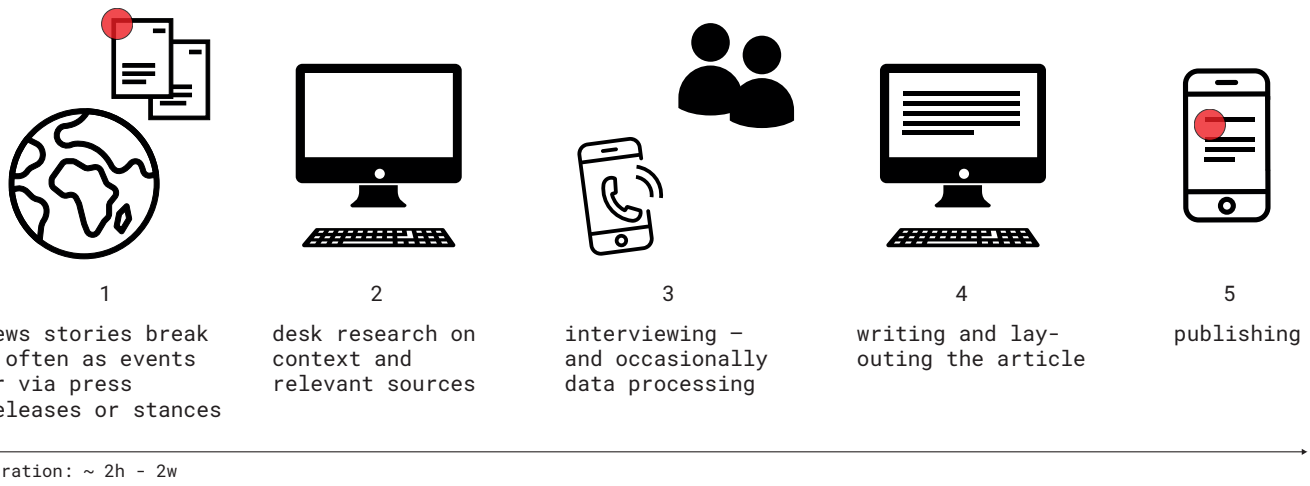


Fig.08 The process of creating a news article as described by the eight participants

The research process (phase two in Fig.08 above) is mentioned by five out of eight participants as especially time consuming. Six out of eight participants further mention that they spend a disproportionate and frustrating amount of time **trying to find the right person (source) to comment on a topic of interest (phase three in Fig.08 above).**
Participants mention that experts, professors, analysts, and politicians are often re-used across articles on a specific topic, thus journalists often re-search in related articles to find relevant sources. Such searches are however difficult to perform on the traditional Web and within a media's own archive of articles, as persons are not semantically related to key terms or organisations. Furthermore, these persons' contact details can be time consuming to find.

> (...) sometimes it takes half an hour to find the right person and his phone number. That's why I find it ridiculous that we do not have a shared list with relevant people and their contacts (...)
> **(Participant 04)**

> (...) I've spent a ridiculous amount of time trying to find the resident chairman in Vollsmose. The area is divided into parks, and the chairman of Bøgeparken is called something and has changed his number... It can be very frustrating.
> **(Participant 01)**

Even though it seems obvious to index used sources and their contact details in some sort of archive, the current practice of sharing contact details and sources is very limited across most media. Only Participant 07 mentions a shared database of relevant sources and contact details for the media's journalists to use. Instead seven out of eight participants have a personal archive consisting of local files, transcripts, and contact details on their computer or smart phone.

> I know some of my colleagues have all sorts of documents with contact details and stuff. But.... When I transcribe my interviews, I write down

> name and number of the person. I can then search for it in Word.
> **(Participant 02)**

Local archives and searches within limited silos as described in the quote above contradicts even with the concept of the traditional Web. Instead, it is recommended to develop a semantic database of sources and contact details allowing a media's journalists and editors to quickly identify relevant sources in relation to specific search terms. Contact details are not described within news articles, thus such information must be added and indexed.
In addition to the sources already used, new sources or even external datasets of contact details can be added to the linked database.
This type of application is referred to as Semantic archive of sources and contact details and is the first out of three applications proposed in this study.
The application for a semantic archive of sources and contact details is examined and discussed in technical detail in Sec.5.2.

4.3 Search in news articles

As illustrated in Fig.08 above, second phase in the process of creating news articles consist of desk research. Seven out of eight participants start their research by searching for relevant articles in their own or other media's online archives to get an overview of context, facts, and sources. Often news articles are related to previous actions and events, thus research in already-published articles is meaningful.
Research in other data sources is also mentioned by all of the participants as an important part of their work process. Methods for this type of research and analysis varies from project to project and often includes extraordinary access to documents or complex data analysis, thus it is not processed further in this study.

Journalists and editors are not able to perform complete searches in their own media's archive of online articles when they are researching for related articles. News articles from approximately the past ten years exist online, but they are not indexed or archived in a way that allows for complete search queries on key words, persons, or organisations. According to Participant 05 - digital editor in chief at Berlingske and responsible for the archive of online articles - it is not even possible to query a complete list of all articles published the past five years.

> We have access to the same archive as the users. Of course, we also have Bond[15] our CMS, but the search options there are very limited. Actually, we have closed down the possibilities to perform content search because the archive is now so extended... if you perform too many large searches the database simply breaks down.
> **(Participant 05)**

Instead, participants describe Google and occasionally Infomedia[16] as the best search tools available for finding related articles. According to the participants, Google's PageRank-algorithm[17] is however not suitable for this kind of search. Articles might not be linked to relevant key words, and it can be difficult to get results published several years ago which makes the search process laborious, inconsistent, and highly time-consuming.

Several participants emphasise the importance of knowledge sharing between colleagues and domains to ensure coherence and to speed up the research phase. To compensate for the inconsistent search, information about related articles is often shared as word of mouth on editorial meeting or between close colleagues.

---

15 Bond is an acronym for Berlingske on Drupal (version 7.0). Drupal is a market-leading content management system.

16 Infomedia is a Danish company monitoring content from most national, regional, and local media as well as news agencies.

17 PageRank is used by Google Search to rank web pages in their search engine results and is a way of measuring the importance of website pages. The algorithm is occasionally changed which affects the listing of search results.

> A common memory impacts the quality of journalism a lot. It means that we can build upon what we already know, and that we don't have to start all over again every time.
> **(Participant 01)**

> We could save an incredible amount of time, if we didn't have to start all over again every day.
> **(Participant 05)**

The analysis documents a search practice – shared by all of the participants – which is random and inconsistent when it comes to finding related articles. This proves a need to semantically index the archive of online articles to enable more precise and complete search queries and to empower knowledge sharing and speed up the process of desk research.

At minimum, the analysis finds that semantic annotation of persons, organisations, places, and key terms is required to perform the kind of search needed.

Such annotation is rather extensive an might lead to extra and time-consuming work for journalists and editors. Interestingly, journalists already spend time on applying metadata – in the form of tags and topic pages – for their articles.
All participants – representing four of the largest news media organisations in Denmark – categorise their articles into superficial categories or topic pages such as politics, international, and culture.
Additionally, Berlingske uses predefined keyword tags to index their articles. It is the journalists' responsibility to manually add tags and topic pages. However, most participants – including the editor in chief at Berlingske who is responsible for the tagging process – find the tagging inconsistent and close to useless.

> I just think... I find it retarded [laughs] that we have to manually add tags. It

> should... the text should be analysed, or the process should be automated. **(Participant 04)**

These tags and topic pages are primarily used to index what category gets the most page views. As participants also express, this does not seem to be a valuable way of using the time spend on manual tagging. Instead it is recommended to look into how this process can be improved to support Semantic Web applications. It should be emphasised that journalists are not trained in writing HTML or any annotation syntax, and the annotation process should not require such technical skills.

In Sec.5.3. a Semantic Web application for Internal semantic news article search is examined and discussed as the second of three types of Semantic Web applications.

As earlier mentioned, journalists often research not just in their own but also in other media's archives; thus, it would be many times more valuable if all media agreed on how to index, annotate, and query online articles. Possibilities for shared standards and practices across multiple media organisations is discussed in Sec.5.3.3.

4.4 Providing layers of information

All of the participants (a total of 32 references across all eight transcripts) describe that providing relevant information in a way that creates a sense of proximity and overview for the users is essential when publishing online news articles. The category is especially urgent in PD II where participants were asked to analyse three online articles from a user perspective. As a result of this activity (see Appendix 9.20), seven out of eight participants suggested the

ability to click or hover on key terms, persons or organisations to get more information about these concepts. This was referred to by a number of participants as layers of information**.**

> If our readers asked themselves: Hey, what was that about? Then they could delve into more information about... eh... the corona virus or whatever. We could also avoid describing x, y, and z over and over again.
> **(Participant 05)**

Participants use tools such as infoboxes **to add additional encyclopaedic information. Sometimes these infoboxes are manually copied and reused in other articles about the same topic, this however requires that the journalist is aware of the existing infobox and know where to copy it from.**
Seven out of eight participants express that they do not think that online news articles today deploy the opportunities of providing elaborate information about concepts mentioned in news articles. Most online articles do not differ remarkably from articles published in newspapers: They primarily consist of static text, and only occasionally provide hyperlinks to other sources of information. If the user wants to explore these additional sources, he or she needs to follow the link and leave the article.

All participants emphasise that users should be able to easily skip additional information, and that implementation of this kind of information must not require extra time-consuming research. When discussing this, three out of eight participants argue that news media organisations and encyclopaedias need to be distinguished.

> We are a news media not a database. People do not turn to us to find out... eh, how many members of Parliament have a journalistic education (...) My fear is that we then tend to... to build a database of information just because it is nice to have.
> **(Participant 05)**

Participants also emphasise that journalists first and foremost ought to be critical towards authorities, and that this is more important than to provide factual information and news you can use.

> Interactive elements are great if they add to the story, but if I were to spend a lot of my time making funny widgets to make people wiser on different aspects of the world... then I would no longer consider myself a journalist, I think.
> **(Participant 03)**

It can be concluded that there is a need for better and faster implementation of additional, encyclopaedic information about persons, organisations, locations and key terms in news articles. It should also be stressed that journalists are expected to focus their resources on investigative research and not spend more time than necessary on encyclopaedic infoboxes.

These requirements can be met by opportunities for reusing and autogenerating information within the concept of Semantic Web. It is crucial that autogenerated content is firmly fact-checked and up to date: If a single infobox contains incorrect information, users might question the reliability of all other content on the media's platform.
Instead of developing large-scale systems that can generate all sorts of infoboxes, it is recommended to focus on a specific type of information to ensure the quality of information.

Participants stress that small summaries of persons mentioned in news articles is a type of infobox frequently used. This type of information is generic and could ultimately by autogenerated.
Within the concept of Semantic Web this is possible if information about a specific person – this could be a media's previous descriptions of that person – is semantically annotated. This type of Semantic Web application enables reuse of a media's content and ensures trustworthiness as the application re-

lies on fact-checked information from the media's own archive.

Participants do not agree on whether a news article in itself – without additional information such as autogenerated summaries – should contain all relevant information, or if it can be expected that users look up information on their own. This disagreement can be interpreted as a movement away from traditional article-based reporting towards more platform-centred journalism, and it is recommended for news media to discuss and examine in more detail when developing applications for autogenerated content.

Development and implementation of the third type of Semantic Web application: Semantic Infobox: Summary for autogenerated summaries is examined and discussed in Sec.5.4.

4.5 Trust and reliability

A final major category identified during analysis is Trust and reliability which is mentioned as crucial to news journalism by all participants (a total of 51 references across all eight transcripts). This category is two-sided: First and foremost (38 references), participants discuss fact-checking and reliability of the sources they use when creating news articles (as part of phase two and three in Fig.08). The other aspect of the category is about protecting and reinforcing the media's trustworthiness and reliability in the eyes of its users. This has been emphasised in recent years as the phenomena and discussion on Fake News has increased.

In the process of creating news stories, five out of eight participants primarily use experts such as professors or analysts to validate information and causalities, and to comment on correlations found in datasets.
Sometimes these experts are also used instead of data to bypass time con-

suming research and analysis. In such situations, journalists are not able to check documentation and argumentation but have to rely solely on the expert's objectivity and reliability.

> It is much easier to call someone who thinks to remember something which makes him an expert. Then you have to rely on that, and then we can write the article, move on.
> **(Participant 01)**

Participants describe this as an accepted condition even though a majority admit that it is not best practice, and that they prefer to get access to the construction of arguments and stances.

> It is impossible to do any kind of reverse engineering when you do not have the documentation. I think it is totally frustrating.
> **(Participant 04)**

Interestingly, some of the participants use the same argument when presented with the concept of linking different online data sets with the support of Semantic Web technologies. Participants find the method powerful but are sceptical towards a potential lack of transparency.

> A good thing about doing the analysis myself is that I think I get a better sense... that I am in control and somehow responsible (...) It is reassuring to have seen the calculation with my own eyes instead of just getting the answer.
> **(Participant 07)**

All participants represent independent media organisation[18] and highly believe that they themselves influence and contribute to the protection and enforcement of their media's trustworthiness.

---

18 Media's independency has often been contested as a normative principle in media policy and journalism where each journalist should make decisions and act according to her own logic. The term independent media is often used to distinguish media organisations from state media (Karppinen, K. & Moe, H. 2016).

Unsolicited, seven out of eight participants describe how Semantic Web technologies can potentially strengthen the reliability of the media they work for. As an example, if additional information about an expert's educational background, his scientific articles, and seniority is easily provided in relation to a news article (as described in Sec.4.4), this would enforce transparency and reliability of that expert. Deduced, this also improves trustworthiness of the media which is transparent about its use of expert sources[19].

It can be concluded that documentation and transparency with calculations and causalities are crucial when designing Semantic Web applications in the context of news journalism. Trustworthiness and proof of documentation is however an Achilles' heel for the concept of Semantic Web (see Sec.3.6). This is a huge – though limited described – disadvantage for Semantic Web technologies in the context of news journalism as applications are of no use if proof and transparency cannot be guaranteed.

However, news journalism as a genre contains qualities to overcome this challenge: News articles are per definition well-researched, and the media and its journalists per definition guarantee that facts and statements are documented and trustworthy. Semantic Web applications relying exclusively on information from news articles (the media's own database) constitute a closed but trustworthy Semantic Web. When participants were asked to come up with hypothetical examples of search queries for a Semantic Web, 10 out of 22 queries (see Appendix 9.15) relied on information from a media's own archive of articles (internal RDF links).

A closed graph – of internal RDF links only – is however in conflict with the core concept of Semantic Web as an open global graph combining different

19 News media organisations' (mis)use of biased experts who pretend to objectively comment on current event is central to the discussion about Fake News.

domains of information. If all organisations designed Semantic Web applications to rely only on data from their own databases, then there would simply be no Semantic Web.
As a starting point it might be fruitful to design Semantic Web applications relying exclusively on internal RDF links as this allows media organisations to build small-scale but valuable applications that can be implemented and applied immediately. At the same time, these small-scale applications contribute to prepare IT-architecture and infrastructure to take advantage of full-scale Semantic Web applications as soon as systems or certifications for trustworthiness and transparency are developed.
It is out of the scope of this study to document how many Danish linked data providers exist, but linked data is not a standard format among established Danish data providers such as public organisations. Thus this is another reason for developing Semantic Web applications which as a starting point rely exclusively on internal RDF links.

The concept of Semantic Web is designed to be expanded continuously. As described in Sec.3.3, RDF graphs can quite easily be merged, thus it is possible to build an application relying on internal RDF links with the purpose of expanding that later to include external sources too. This is discussed further in Sec.5.2.4, 5.3.3, and 5.4.4.

4.6 Partial conclusion (answering RQ1 + RQ2)

The qualitative analysis reveals three areas within Danish news journalism with significant potential of improvement in the context of Semantic Web.

The first area concerns the challenge of finding the right person to comment on or evaluate a specific topic. Journalists often research in related articles to find

relevant sources. Such searches are however difficult to perform on the traditional Web as persons are not semantically related to key terms or organisations. This part of the 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 contact details 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 and is discussed in technical detail in Sec.5.2.

The second area concerns issues of finding previously published articles related to a specific concept. When news break, journalists often search for context, facts, and sources in related news articles published by their own media or by others. Journalists and editors are however not able to perform complete searches even in the media's own archive of articles. Journalists experience inconsistency and limitations in standard search engines such as Google where it can be difficult to search for articles published a long time ago.
This process can be empowered through semantic annotation: If all persons, organisations, places and key terms described in a media's archive of articles where semantically annotated, it would be possible to perform thorough and complex search queries. Such annotation is laborious and time-consuming, however the analysis reveals that journalists already spend time on tagging and annotating articles before publication. This tagging process is superficial and does not support searchability remarkably, and it is recommended to look into how it can be improved to support Semantic Web applications instead. Sec.5.3 discusses in more technical detail how a Semantic Web application for Internal semantic news article search can be implemented.

The third area targets improvement of user experience 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 about persons or other concepts mentioned in news articles. On the other hand, it is demonstrated that such additional information contributes to transparency and ultimately supports a media's trustworthiness. Especially infoboxes about experts or other authorities can be used as documentation for why these are chosen as expert sources.
The analysis indicates that Semantic Web technologies can be applied to autogenerate this type of additional information. 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 and is discussed in technical detail in Sec.5.4.

For all of the three areas of interest and the suggested solutions, the importance of documentation, proof, and transparency is emphasised. This is an Achilles' heel for the concept of Semantic Web as standards or certifications to guarantee trustworthiness are not yet developed. To overcome this challenge, the suggested solutions can be designed to rely exclusively on a media's own trusted sources of information (internal RDF links) as a starting point. This is however in conflict with the core concept of Semantic Web, and it is recommended to extend the applications in a controlled way to also include external sources of linked data.
Opportunities and challenges of such expansions are discussed in Sec.5.2.4, 5.3.3, and 5.4.4.

5.0 Technical analysis and discussion: Creating news journalism in the context of Semantic Web

Chapter 5 discusses how three types of Semantic Web applications can be developed to support the needs of Danish news journalists identified in Chapter 4.0.
Sec.5.1 outlines the three types of applications, before each of them is described and discussed in more detail in Sec.5.2–5.4. These sections focus on what technical requirements needed for each application to be realised. The chapter concludes with a partial conclusion in Sec.5.5 answering RQ3.

5.1 Developing Semantic Web technologies for news journalists

Analysis of the qualitative interviews and PD studies with journalist participants found several areas where the work process of Danish news journalists can be improved. The areas span from the process of research and finding contact details to elements of improved user experience (see Sec.4.1–4.5); however, the areas also have common traits: They pivot around optimising the amount of time that journalists need to spend on specific tasks (see Sec.4.2-4.3) and at the same time ensuring trustworthiness and reliability (see Sec.4.4-4.5).

On basis of this analysis, three specific applications are proposed in this study (an introduction of each application can be found in Sec.4.6):

- Application I: Semantic archive of sources and contact details
- Application II: Internal semantic news article search
- Application III: Semantic infobox: Summary

The applications match the concept of Semantic Web anno 2020 and accede to the most common needs identified in the qualitative research.
It is important to note, that the three applications do not comprise a complete list of how Semantic Web technologies can support Danish news journalism today or in the future. It is possible to imagine more powerful solutions; however, these transcend the scope of this study and also the technological and practical limitations of 2020.
The aim of this study is to identify Semantic Web applications with documented value to journalists working in Danish newsrooms today and technically examine and discuss how these can realistically be implemented.

5.2 Semantic archive of sources and contact details

The following sections contain step-by-step examination of how a semantic archive of sources and contact details can be realised. The aim of this type of application is to allow a media's journalists and editors to quickly identify relevant sources and their contact details in relation to specific search terms. The section concludes with a discussion on challenges and perspectives on further development.

This type of application raises questions about privacy and GDPR regulations as people's names are connected and archived together with personal information such as phone numbers and e-mail addresses. It is out of the scope of this study to go into a detailed discussion about legal rights and ethics as these topics are extensive enough for a separate examination. It should however be mentioned, that the application is thought of as an integrated part of the media's CMS or intranet and thereby protected by a personal login authorisation process. This means that only the media's journalists and editors can

access the information, and that an administrator has the possibility to avert specific types of data combinations or exclude some user groups from parts of the contained information. This way, the application can be modified to cater for regional regulations or work ethics.

5.2.1 RDF graphs, URIs and vocabularies for sources and contact details

As mentioned in Sec.3.3, RDF graphs are used to visualise how statements can be expressed using subject-predicate-object triples. Fig.09 below shows the simplest graph for describing sources in news articles. This is the minimum amount of information needed to build a functional archive of sources and contact details where journalists can easily evaluate the 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 (see Sec.3.2).

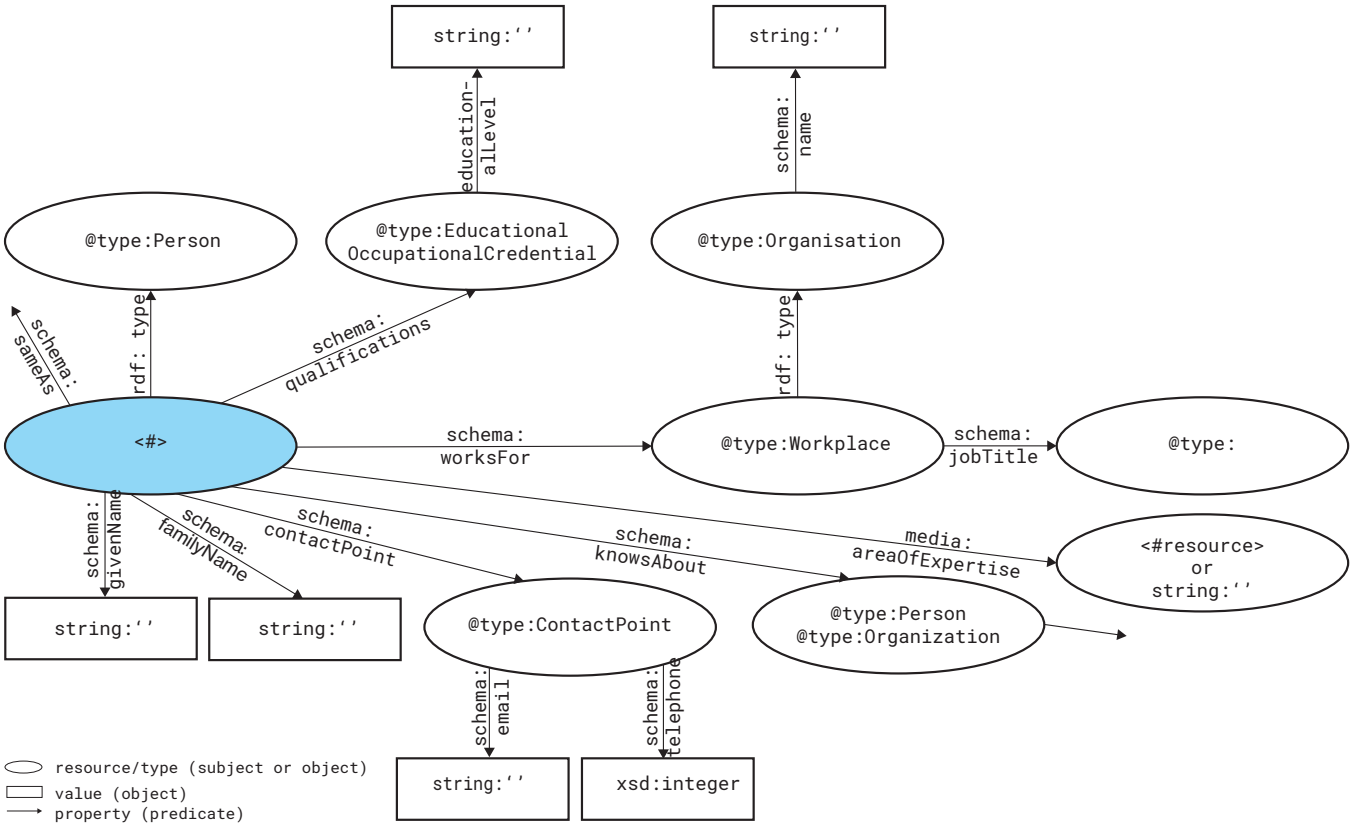


Fig.09 Minimum graph structure for semantic description of sources and contact details. In this example, namespaces are used in instead of URIs to improve readability.

All resources (represented by circles in Fig.09 above) must be organised in a formalised ontology with class hierarchies and domain- and range relations (see Sec.4.3). Taxonomies and ontology rules can be created using the free, open source editor Protégé. This is however extremely time consuming, thus it is recommended to use existing standard vocabularies to describe RDF-triples as URIs for resources and properties as well as the underlying taxonomies and rules of the standard vocabulary can then be reused.

The generic example in Fig.09 above demonstrates how the standard vocabulary Schema.org[20] can be used to describe a given source with his or her con-

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

tact details as well as information about workplace, title, level of education etc. Schema.org is however limited when it comes to linking a source to its areas of interest or expertise. In the example above, the Schema.org-property knowsAbout is partially used to describe this. Schema.org recommends[21] that knowsAbout is used for linking a resource to other Persons or Organisations but not topics such as Epidemic, Infrastructure, or Banking. In order to do so, the vocabulary needs to be extended with a property such as areaOfExpertise.

It is out of the scope of this study to describe in detail how standard vocabularies can be extended, but it is worth briefly mentioning that any extension should be done in controlled namespaces and with terms from RDFS and OWL to relate the new resource or property to terms in an existing vocabulary (Hendler, Heath, & Bizer, 2011, p.63). The construction of new resources or properties should also live up to the Linked Data Principles so that Semantic Web applications can look up their definitions.
All news media organisations examined in this study own at least one domain[22] (namespace), thus these are obvious to use for minting URIs for new properties such as areaOfExpertise, e.g.: http://media.dk/vocab/0.1/areaOfExpertise

Equally, each source (Person) should be named and described with a set of URIs according to the Linked Data Principles (see Sec.3.2) within the media's domain (see generic examples in Table 09 below).

Types of URI-reference	Example
URI for the real-world person	http://media.dk/resource/Person/Soeren-Brostroem01
URI for a related information resource that describes the real-world person and has an HTML representation	http://media.dk/page/Person/Soeren-Brostroem01

21 See recommended use: https://schema.org/knowsAbout
22 In this study, participants represent news media organisations with the following domains: http://www.berlingske.dk, http://www.tv2.dk, http://www.dr.dk, and http://www.jyllands-posten.dk

URI for a related information resource that describes the real-world person and has an RDF/XML representation	http://media.dk/data/Person/Soeren-Brostro-em01

Table 09 Types of URI-references (example)

To fulfil the areaOfExpertise-triple - or any other triple in Fig.09 with a resource as the object - the vocabulary must include or be extended with resources describing each object as illustrated in Fig.10 below.

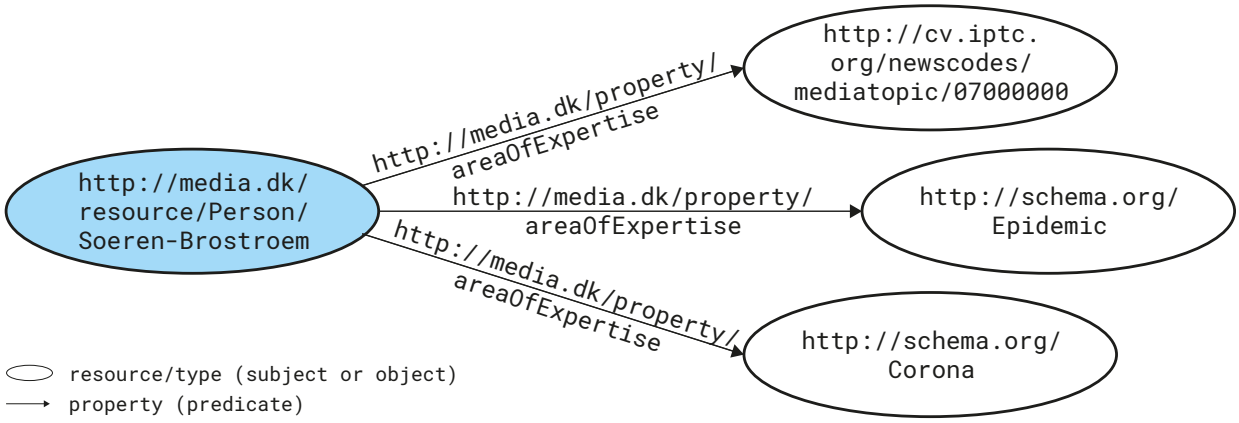


Fig.10 RDF graph showing areaOfExpertise for the source Søren Brostrøm, head of Danish Health Authority

The list of resources needs to be rather extensive, and it is recommended to reuse as many existing resources from standard vocabularies as possible. As a starting point, concepts from The International Press Telecommunications Council's (IPTC) Media Topics NewsCodes[23] can be adopted. This ontology contains hundreds of resources tailored to describe news article content in different languages, including Danish. The use of multiple sets of vocabularies is further examined in Sec.5.3.1.

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 annotati-

23 https://www.iptc.org/std/NewsCodes/mediatopic/treeview/mediatopic-en-GB.html

on as it can be hard to integrate and might cause inconsistency or missing links. To overcome this, sameAs-relations can be used to juxtapose similar terms in different languages (see Fig.11 below). Thus, it is recommended to create new resources in Danish for all things and terms and link these to English resources – if possible, some which already exist in standard vocabularies – via sameAs-relations. This ensures that the Danish resource becomes part of the global graph and allows standardised software to be applied (see Sec.5.2.2).

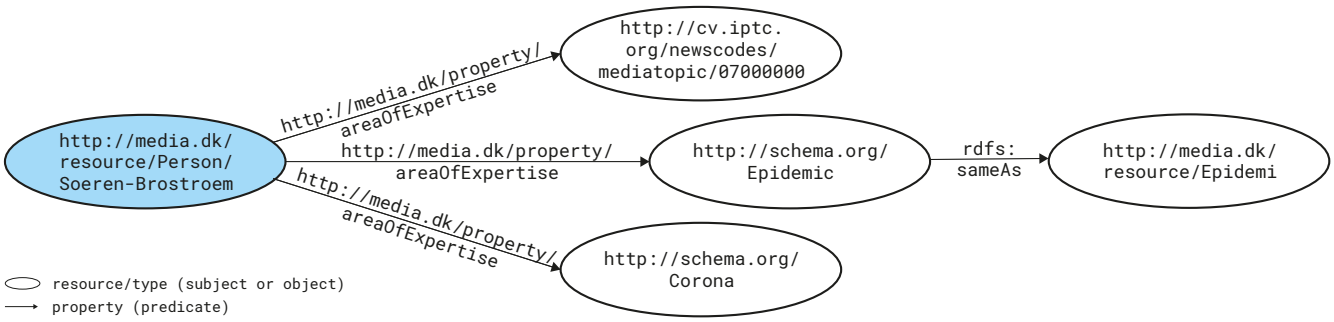


Fig.11 SameAs-relations to juxtapose resources in different languages. The uppermost resource does not need to be 'translated' as this is a global resource from the IPTC-vocabulary. The bottom resource is a name and does not need to be 'translated' either.

Named resources such as Persons and Organisations can be directly reused in English as there most often is no difference in the English and Danish phrasing of these.

For this type of application, properties from standard vocabularies can also be reused in English as these are not being displayed (instead symbols are used, see Fig.14). Annotation and querying in different languages are further discussed in Sec.5.2.3.

Additionally, it is recommended that Danish news media organisations who use tags to index their articles (see Sec.4.3) convert these into resources and apply Linked Data Principles to create URIs for each tag (see Sec.3.2). This way, the indexing which is already applied by tags can be used as a starting

point for the semantic annotation. Each resource representing a tag in Danish should be linked to a resource (from an existing standard vocabulary) describing the same thing but in English as discussed above.

For clarity, additional examples and graphs in this study only include resources in English. Each of these can potentially be linked via sameAs-relations to resources in Danish.

5.2.2 Serialisation and annotation of sources and contact details

Semantic annotation can be implemented in each news article using RDFa as described in Sec.3.3.1 and illustrated in Snippet 04 below. The HTML code is extracted from a news article[24] about Covid-19 and contains semantic information about head of Danish Health Authority, Søren Brostrøm, who is used as a source in the article.

```
<div class="dre-speech" vocab="http://schema.org/" property="about" typeof="Person"
resource="http://media.dk/resource/Person/Soeren-Brostroem01">
    <p class="dre-article-body__paragraph dre-variables">- Der er jo ingen tvivl om, at vi nu har det, vi
    kalder samfundssmitte i Danmark, siger<span property="worksFor">
        <span vocab="http://schema.org/" typeOf="Organization">
            <span property="name">Sundhedsstyrelsen</span>s
        </span>
        </span>
        <span property="jobTitle">direktør</span>, <span property="givenName">Søren </span><span
        property="familyName">Brostrøm</span>.</p><span property="contactPoint">
        <span vocab="http://schema.org/" typeof="ContactPoint">
            <meta property="email" content="sb@sund.dk"/>
            <meta property="telephone" content="+4551283028"/>
        </span>
        </span>
        <span property="qualifications">
        <span vocab="http://schema.org/" typeof="EducationalOppucationalCredential">
            <meta property="educationalLevel" content="professor"/>
        </span>
        </span>
    <p class="dre-article-body__paragraph dre-variables">Tallene bekræfter nemlig, at smitten nu florerer i
    det danske samfund. Og det er et ganske andet billede, end da <span property="areaOfExpertise"><span
    vocab="http://media.dk/" typeOf="Topic"><span property="name">corona</span></span></span>virusset først
    gjorde sit indtog i landet for et par uger siden.</p>
</div>
```

Snippet 04 HTML with RDFa annotation of source details. Note how information about contact details are added as meta-tags which allow them not to be displayed in the published document.

24 https://www.dr.dk/nyheder/indland/vi-har-nu-det-man-kalder-samfundssmitte-i-danmark

Based on this annotation, an RDF-description of the source Soeren-Brostroem01 can be generated and viewed by looking up the URI http://media.dk/data/Person/Soeren-Brostroem01 (see Snippet 05 below). This description constitutes data points for a semantic archive of sources and contact details.

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF
   xmlns:ns1="http://www.w3.org/ns/rdfa#"
   xmlns:ns2="http://schema.org/"
   xmlns:ns3="http://media.dk/"
   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
>

  <rdf:Description rdf:about="http://media.dk/resource/Person/Soeren-Brostroem01">
    <ns2:givenName>Søren</ns2:givenName>
    <ns2:familyName>Brostrøm</ns2:familyName>
    <ns2:worksFor>Sundhedsstyrelsens</ns2:worksFor>
    <ns2:jobTitle>direktør</ns2:jobTitle>
    <ns1:qualifications>professor</ns1:qualifications>
    <ns2:contactPoint>sb@sund.dk</ns2:contactPoint>
    <ns2:contactPoint>+4551283028</ns2:contactPoint>
    <ns2:areaOfExpertise>corona</ns2:areaOfExpertise>
    <rdf:type rdf:resource="http://schema.org/Person"/>
  </rdf:Description>
</rdf:RDF>
```

Snippet 05 RDF/XML description of the source http://media.dk/resource/Person/Soeren-Brostroem01

To ensure transparency and traceability, it is recommended to add metadata to the RDF/XML description about the person who has applied the semantic annotation and how recently the dataset was updated (see Sec.3.6 about Semantic Sitemaps).

RDFa annotation in Snippet 05 above is manually added. This is 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 this annotation.

In 2008, Thomson Reuters launched the linked data entity extractor Calais (see Sec.1.1.1) 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, p.35). Today, Calais is integrated in the latest version of Drupal which is the content management system (CMS) used by Berlingske, Danmarks Radio, and Jyllands-Posten. This study will not go into detail on how Calais can be activated in Drupal and applied in a Danish context as this is an area extensive enough to be a study on its own. Also, it requires an update to the latest version of Drupal which neither Berlingske, Danmarks Radio, or Jyllands-Posten has today.

Larger news media organisations such as BBC and N.Y.T. have developed their own annotation systems. BBC Juicer was built in 2011 as an API to extract concepts from news articles and link them to DBpedia resources (see Sec.1.1.1). The annotation is performed completely automatically as an algorithmic process and is not manually or editorially controlled (BBC Newslab, 2018). According to the BBC News Lab, the API was shut down in 2018:

> Since we failed to find any other sustainable uses, we shut it (BBC Juicer API, ed.) down in 2018
> (BBC Newslab, 2018)

It requires dedicated research to conclude exactly why BBC failed to find any sustainable uses. With the technology available today, it is however hard to imagine how quality annotation can be secured in a fully automated process, and one reason can be that the BBC Juicer annotation was not detailed and precise enough to use in truly valuable applications.

In 2015, N.Y.T. launched a similar service as an experimental AI project. This service is simply called Editor and is a semi-automated tool for annotating news articles with semantic information (see Fig.12 below). According to N.Y.T., the application 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 (N.Y.T. Labs, 2015). When journalists write a specific name or

keyword, they are immediately provided with drop-down suggestions for how the concept can be described semantically. This way, annotation becomes an integrated part of the writing process, and the journalist herself possess editorial control of the annotation.

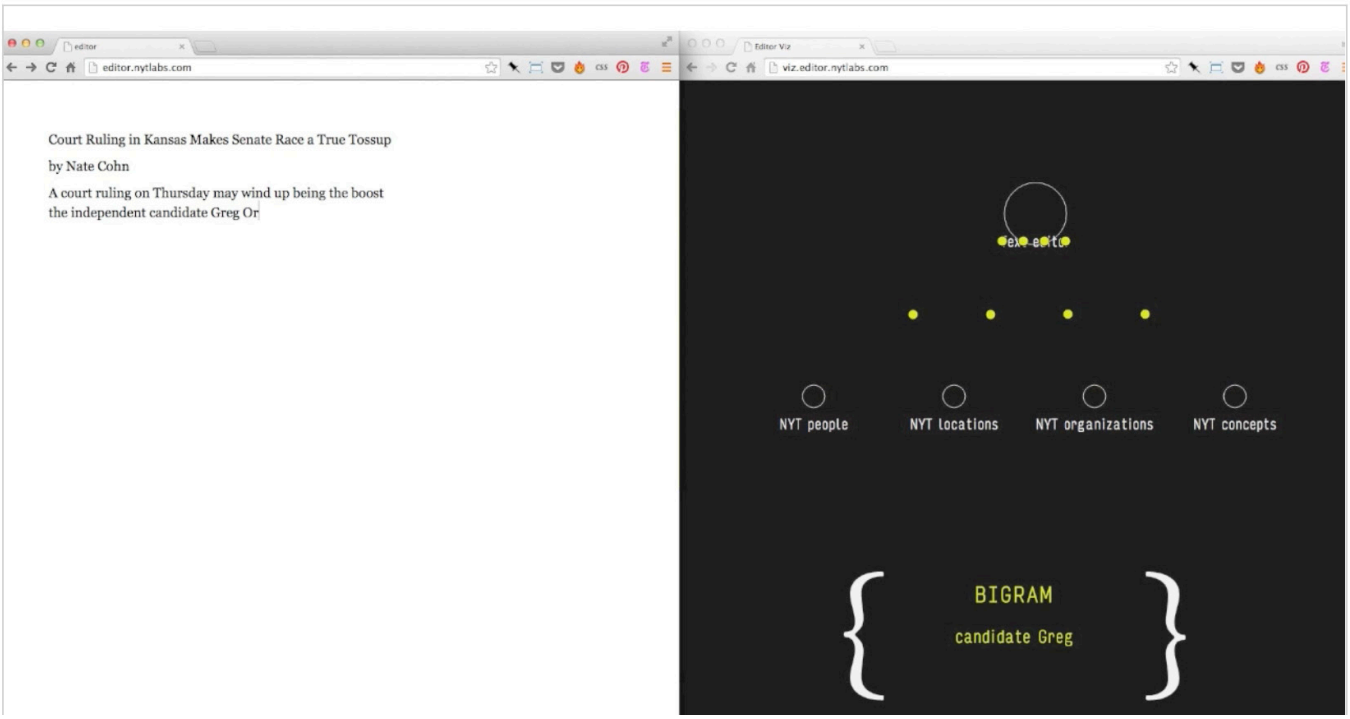


Fig.12 N.Y.T.'s Editor launched in 2015 as an experimental project exploring the collaboration between machine learning systems and journalists.

According to Sandhaus (2012), more than 60 people were involved when N.Y.T.'s Editor was launched. This is more people than work in most Danish newsroom, and the resources put into developing this type of annotation tool is out of the scope of any Danish news media organisation.

Whether the solution is to implement a document annotator like Calais into a media's CMS or develop a tailored annotation system, the practical integration and use of semi-automated annotation systems can be considered one of the most challenging aspect of developing Semantic Web applications for news media organisations today.

Annotation for a semantic application to query relevant sources and contact details can however be kept relatively simple (see Fig.08). Detailed descriptions and multiple relations connected to each resource will improve the search tool and make it more precise, but the application can be limited to a fixed set of properties from only two vocabularies and still be highly valuable to Danish news journalists. This means, that annotation for this type of application does not need to rely on AI but can be developed using existing techniques for traditional search applications.

5.2.3 Information query and user interface

Once semantic annotation is applied, it is possible to query sources and their contact details related to specific organisations, persons or topics.
This can be done using SPARQL-queries and the library SPARQL Lib as illustrated in Snippet 06 below. The example retrieves information from the open database DBpedia (l.81 in Snippet 06) using the endpoint http://dbpedia.org/sparql. Theoretically this endpoint can be replaced by a media's linked database instead.

```
79              <?php
80                require_once('sparqllib.php');
81                $db = sparql_connect('http://dbpedia.org/sparql');
82                $query = "
83
84                PREFIX dbo:<http://dbpedia.org/ontology/>
85                PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
86                PREFIX dbp:<http://dbpedia.org/property/>
87                PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
88
89                SELECT ?name ?title ?workPlace ?education ?phone ?email
90                WHERE{
91                    ?source a dbo:Person.
92                    ?source dbo:knownFor dbr:Epidemic.
93                    ?source rdfs:label ?name.
94                    ?source dbo:occupation ?title.
95                    ?source dbp:workplace ?workPlace.
96                    ?source dbp:education ?education.
97                    ?source dbp:phone ?phone.
98                    ?source dbp:email ?email.
99                OPTIONAL {?phone ?email}.
100               FILTER (lang(?name)="da").
101                   }
102             ";
```

Snippet 06 Example of SPARQL-query requesting information about a relevant source linked to the topic Epidemic through the DBpedia-property knownFor.

In the example, a query is performed requesting name, title, work organisation, education, and contact details for all sources connected to the term Epidemic via the DBpedia property knownFor (l. 92). This is attained using a SELECT clause to list the requested information (l.89 in Snippet 06) and different WHERE clauses to state what relations the information should exist in (l.90–98 in Snippet 06). The OPTIONAL clause (l.99 in Snippet 06) makes sure, that the query does not break, if information about phone or email is not annotated. Finally, the FILTER clause (l.100 in Snippet 06) ensures that results are retrieved in Danish.

For brevity, a set of PREFIX clauses (l.84–87 in Snippet 06) are used to replace URIs with namespaces, e.g. dbo: is used as a PREFIX referring to the DBpedia ontology which the property knownFor is a part of. Theoretically, this can be replaced with a property within the media's own ontology, such as: http://media.dk/vocab/0.1/areaOfExpertise.

Journalists are not trained in writing SPARQL-queries, and it is recommended to develop a search panel and design a user interface to guide the construction and formatting of these queries. Fig.13 below illustrates how such a simple search panel can be constructed.

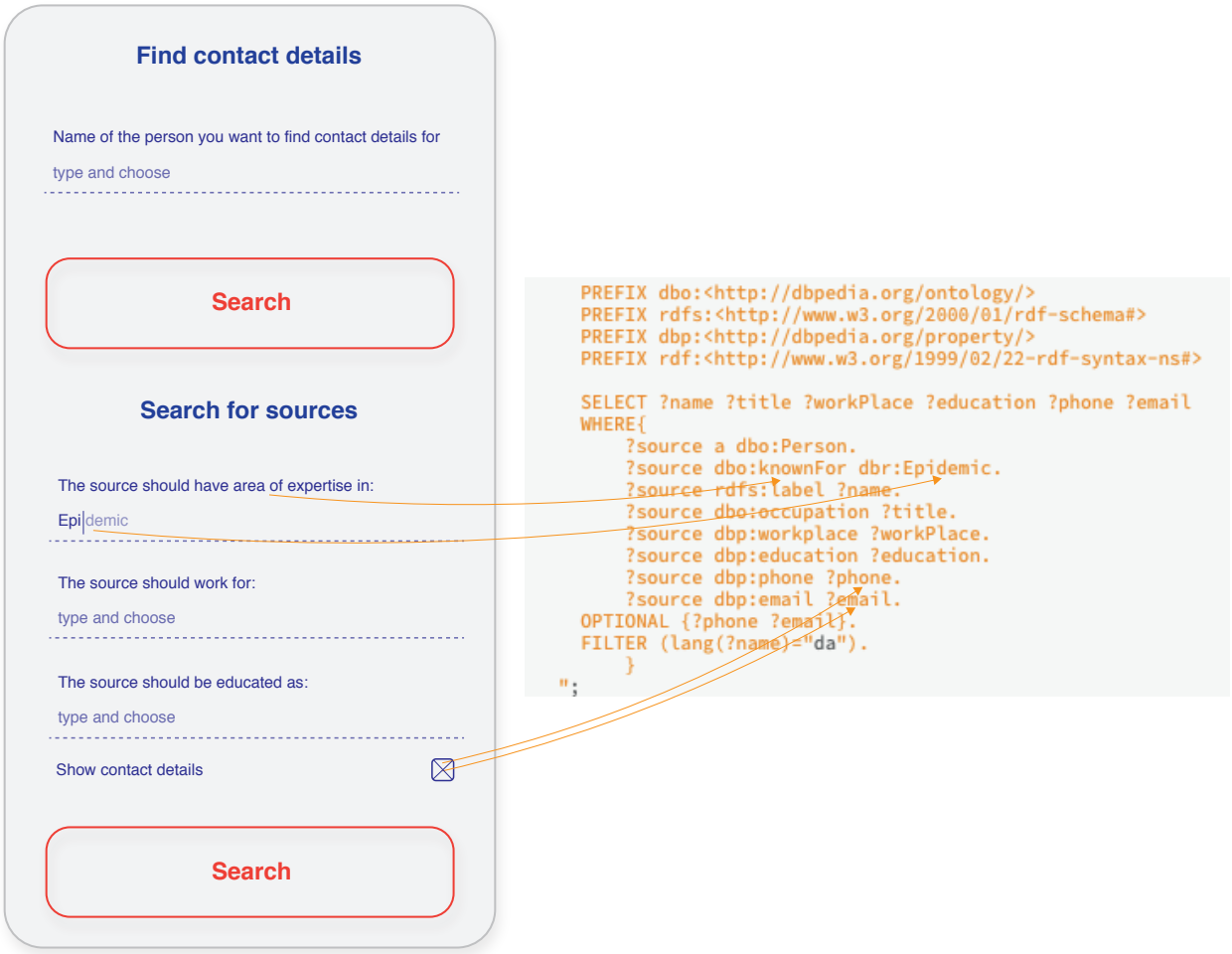


Fig.13 Suggested search panel for sources and contact details search

The search panel includes options to search for contact details of an already known source and to search for relevant sources related to a specific term, a workplace or with a specific educational background.

When searching for sources, 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 the media's controlled vocabulary. These appear from a drop-down list as the user starts typing. The user can fill out all of the parameters or only one of them. The user interface ensures that the combination of properties and values are mat-

ched and formatted correctly following the SPARQL syntax. The technical implementation of this formatting can be performed using the library SPARQL Lib, however it is out of the scope of this examination to go into further detail with this.

The user interface is phrased in English even though the search panel is designed for Danish news journalists. The reason for this is to urge or encourage users to write search terms in English. As discussed in Sec.5.2.1, standard annotation is by default written in English, and resources describing terms in Danish should be linked to English equivalents via sameAs-relations. This enables search queries with search terms phrased in Danish to be performed, however English will always be a more robust search language. This language differentiation weakens usability for Danish news journalists, but annotation in English ensures that RDF triples can easily be linked to international sources. Furthermore, querying in English makes news journalists aware that they are querying a database which is potentially connected to the entire world – this might even encourage them to research for stories they normally would not have thought of.

Information retrieved in the query can be integrated and formatted using PHP: Hypertext Preprocessor (PHP) and different if-statements as illustrated in Snippet 07 below. If the query contains any results, h2-tags with names of the relevant sources are displayed together with information about title, workplace, education, and contact details of each source (l. 111-137).

```php
if($result = sparql_query($query)){
  echo "<h2>Suggested sources</h2>";
  $fields = sparql_field_array($result);
  while($row = sparql_fetch_array($result)){
      echo "<div>";
    foreach($fields as $field)
    {
      $str = $row[$field];
      $str = preg_replace('#^http://dbpedia.org/resource/#', '', $str);
      $str = preg_replace('#_#', ' ', $str);
      if ($field == "name") {
          echo "<h3>" . $str . "</h3>";
      }
          else
          if ($field == "title"){
          echo "<b>" . $str . "</b>";
      }
          else
          if ($field == "workPlace"){
          echo "<b>" . $str . "</b>";
      }
          else
          if ($field == "education"){
          echo "<b>" . $str . "</b>";
      }
          else
          if ($field == "phone"){
          echo "<b>" . $str . "</b>";
      }
          else
          if ($field == "email"){
          echo "<b>" . $str . "</b>";
      }
          ?>
```

Snippet 07 PHP to integrate and format information about source and contact details

Fig.14 below illustrates how a user interface can be designed for a search panel and listing of results. Note how results are retrieved and displayed with values phrased in Danish to increase usability.
Additionally, it is recommended to implement an edit function enabling users to easily edit information if these are not updated. Updated information should automatically be integrated in the resource's RDF/XML description (see Snippet 05 and Sec.5.2.1).

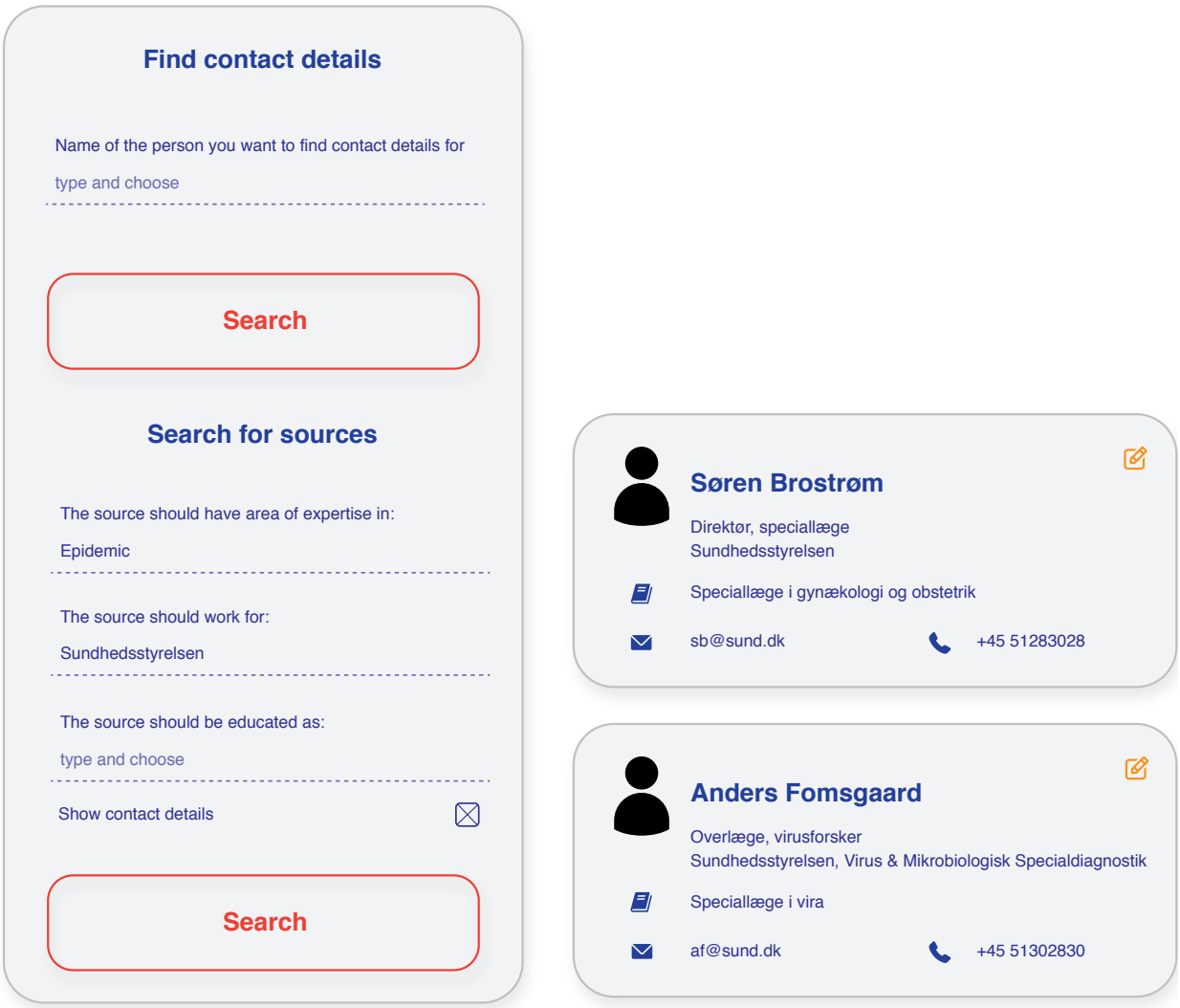


Fig.14 Search panel and results list for sources and contact details

5.2.4 Sources and contact details: Challenges and further development

As mentioned in Sec.5.2.2, practical implementation of semantic annotation is one of the most challenging part of this type of Semantic Web application. First, all the media's tags need to be converted to linked data resources with URIs and RDF/XML descriptions, then a large number of sources and concepts within each news article need to be annotated with semantic mark-up refer-

ring back to the linked data resources.

To reach a critical number of annotated concepts, the established Semantic Web community[25] emphasises large-scale cooperation and user contributions. Same method needs to be applied within a news media organisation: Every journalist must contribute with pieces of semantic information and add this to the database when she comes across new or updated data.
This also means, that the type of application is not suitable for smaller organisations. In fact, it is recommended that multiple Danish news media organisations – and their journalists – join forces and build a database of sources and contact details together. The information contained is generic and not protected by copyrights, and multiple media organisations can possibly contribute to and value from a shared database.

The application – as described in the sections above – rely exclusively on information from a media's own database of articles (internal RDF links). If persons described in all articles published within the past five years were to be annotated, the archive would contain hundreds of relevant sources, and the application would be highly valuable to journalists and editors.
However, if the application truly is to be considered part of the Semantic Web, it should rely on external sources of linked data too (external RDF links). This expansion must be done in a controlled way, where only reliable datasets are included (how to guarantee trustworthiness is further discussed in Sec.5.4.2). Datasets containing name, title, and contact details of all active politicians in Denmark could be retrieved from political organisations or municipalities and linked to the media's database. Ultimately, these datasets consist of already annotated linked data, and the integration would consist merely of linking each source to relevant concepts (in the media's database) through areaOfExpertise-relations.

25 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.

More controversially, a semantic archive of sources and contact details could be linked to a media organisation's e-mail system. Meaning, that every time a journalist receives an e-mail – which most often contains information about the sender's name, title, workplace, and contact details – the sender would be annotated as a possible source within the semantic database. This addition increases the need for further examination of privacy and GDPR regulations and the protection of this type of information in the context of Semantic Web. It is out of the scope of this study to go into such legal discussions, but it is worth mentioning that a solution could be to exclude parts of the media's database from being part of the global Semantic Web ecosystem (see also Sec.5.4.4 for reflections on reliability certifications and Semantic Sitemaps).

5.3 Internal semantic news article search

The following sections contain stepwise examination and discussion of how an application for sematic search for news articles within a media's database can be implemented. The aim of this type of application is to allow a media's journalists and editors to perform complex and complete searches for articles related to relevant concepts within the media's own archive. The section concludes with a discussion of challenges and perspectives on further development.

5.3.1 RDF graphs and vocabularies for semantic news article search

Semantic search for news articles related to specific topics, persons, or organisations requires concepts to be annotated in RDF triples the same way as for the semantic archive of sources and contact details (see Sec.5.2.1).
In order for the search to be detailed and extensive enough, it is found in the

analysis (see Sec.4.3) that all persons, organisations, locations, and key terms described in each news article need to be semantically annotated. Every text paragraph in every news article contains multiple RDF triples describing specific concepts and their relations to other resources. As an example, this allows journalists to query a complete list of all news articles about the Olympic Games 2020, which is not possible with standard search engines (see Sec.4.3). Further, a semantic search application can sort the list of results by date of publication or filter it to only show articles also related to the resource Tokyo.

Fig.15 below illustrates how a paragraph in a random news article[26] can be modelled using RDF graphs. For clarity, the model is simplified and contains blank nodes (see Sec.3.3). In theory these blank nodes can connect the graph to other graphs and combine them to one global graph describing the entire archive of articles.

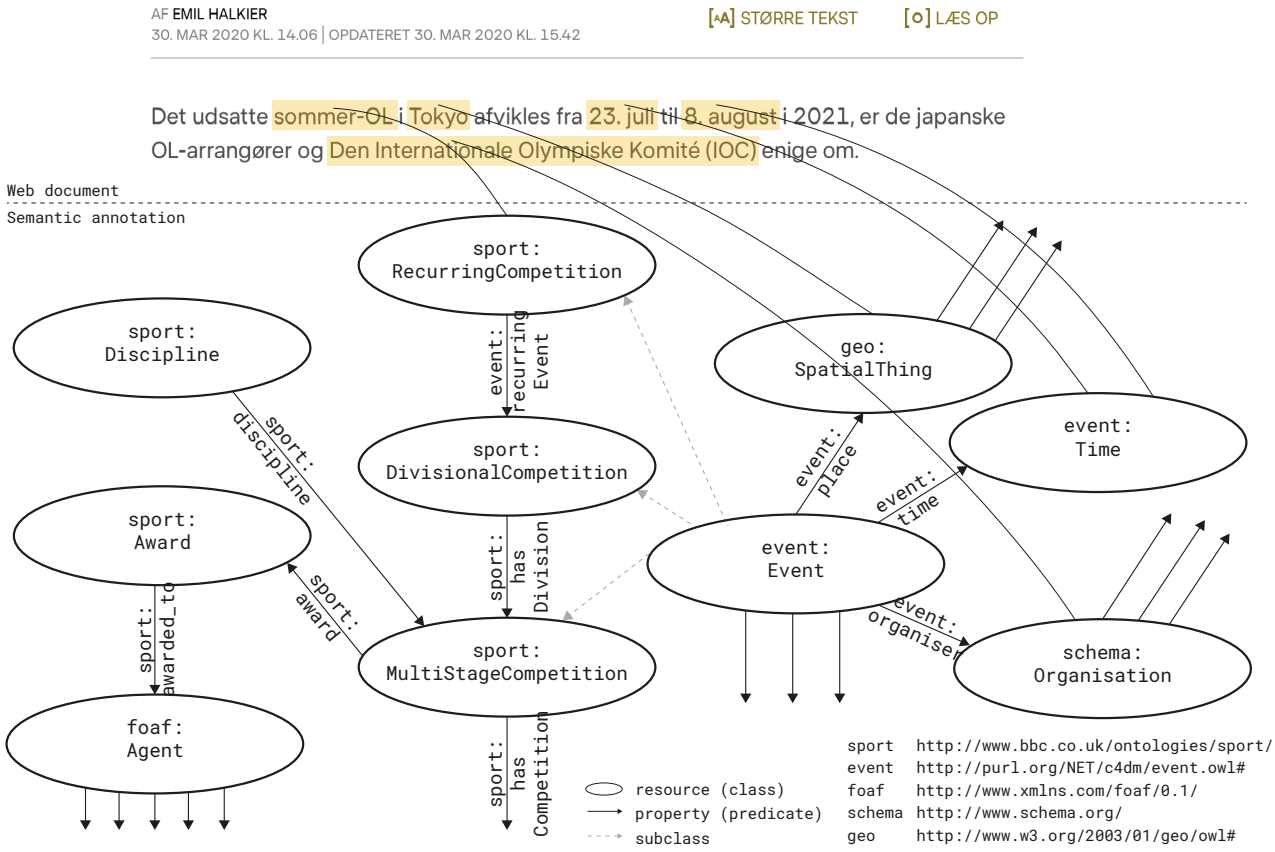


Fig.15 RDF graph for annotating a news article paragraph (simplified for clarity)

26 https://www.dr.dk/sporten/ol/datoer-udsat-ol-er-fundet-legene-begynder-23-juli-2021

This annotation is much more comprehensive than what was needed for the archive of sources and contact details and requires use of multiple vocabularies to describe relevant relations between different domains. In the example above, four of the most common vocabularies (see Table 10 below) are used together with BBC's sport vocabulary designed to describe sport events.

Use of multiple vocabularies constitutes challenges of selecting the most relevant vocabularies and making sure that these are not used simultaneously to describe resources inconsistently. As an example, the vocabularies Schema.org and FOAF contains almost the same properties for describing persons and their relationship to other persons and organisations, thus it should be decided which of the two to use. A selection of vocabularies needs to be firmly integrated in a document annotator (see Sec.5.2.2), and the annotator should be programmed to always have one dominant vocabulary so that similar annotation options (from different vocabularies) are not suggested for describing the same thing.

Vocabulary (ontology)	Domain
http://purl.org/NET/c4dm/event.owl#	designed to semantically describe events
http://xmlns.com/foaf/0.1/	designed to semantically describe persons and their relation to other persons
http:/www.schema.org/	consists of multiple sets of vocabularies and can be used to describe a variety of domains from persons to medical equipment
http://www.w3.org/2003/01/geo/owl#	designed to semantically describe geographical locations
http://www.bbc.co.uk/ontologies/sport	designed to semantically describe sport events

Table 10 Vocabularies and domains

All resources and properties are described with URIs – some of them reused from existing vocabularies, some of them created as new resources within the

media's controlled domain (see Sec5.2.1). As an example, it is recommended to create URI-references within the media's controlled domain for Tokyo and Den Internationale Olympiske Komité (IOC) in the example illustrated in Fig.15 above (see Table 09 for examples of how such URI can be minted). Resources from different vocabularies that describe the same thing can be interlinked through sameAs-relations as illustrated in Fig.16 below. Here the country of Denmark is represented by one URI-resource in N.Y.T.'s vocabulary and another URI-resource in the DBpedia ontology.

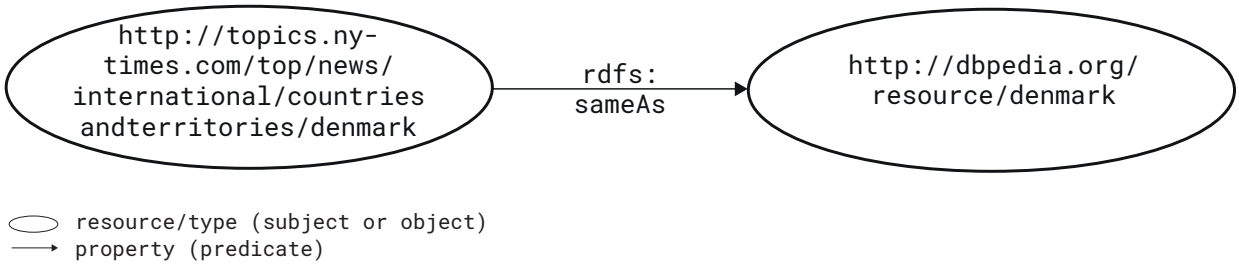


Fig.16 sameAs-relation showed as RDF graph

This interlinking ensures that the resource in N.Y.T.'s vocabulary inherits the rules and taxonomies (see Sec.5.2.1) of the resource in DBpedia.

As described in Sec.3.3.1, it is important to distinguish real-world objects from the HTML document that describe these (the news article). Thus, it is recommended to annotate not only concepts and domains but also metadata about the Web document and its relation to concepts described within it (Hendler et al., 2011, p.19). Fig.17 below illustrates how such relations can be modelled using RDF graphs.

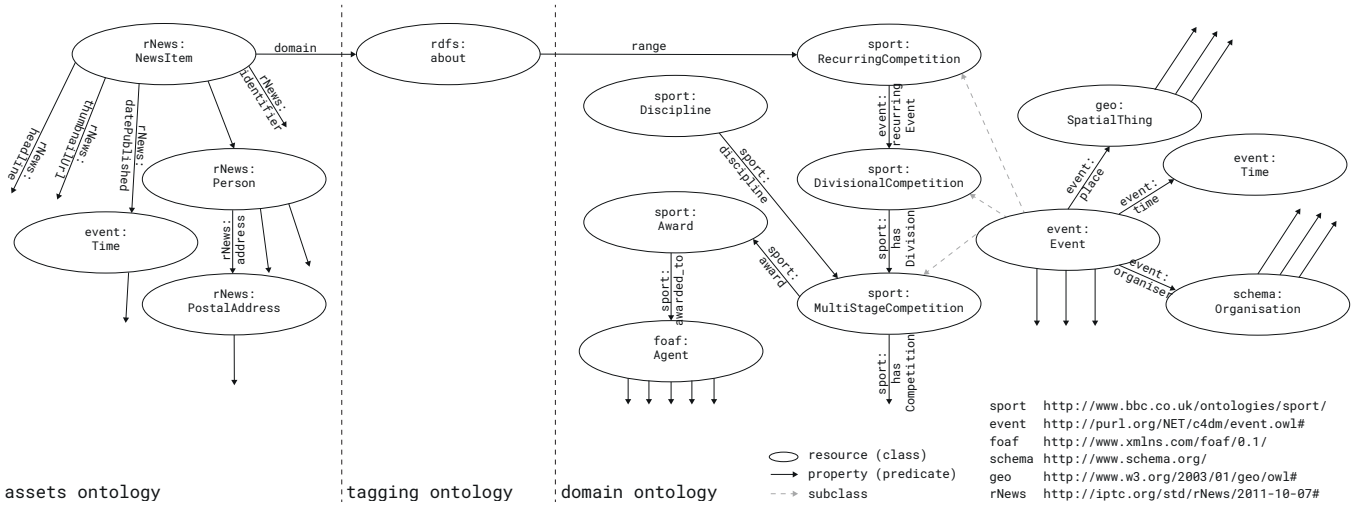


Fig.17 RDF graph for assets-, tagging- and domain ontologies (simplified for clarity). In this example, the rNews-ontology is used to annotate metadata about the NewsItem.

In the example, the rNews ontology[27] is applied to annotate information about author, publication data, and thumbnail-URL. The vocabulary is designed by IPTC to ensure consistency in how news media annotate metadata. Fig.17 above describes what metadata (in the area assets ontology) are recommended to annotate about online news articles. 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 articles, a robust annotation system – integrated into the CMS – is needed (see Sec.5.2.2).
Even with an annotation system such as N.Y.T.'s Editor, the process of semantically describing an entire news article requires more time, than journalists spend on tagging articles today. Current technology still requires manual editing to secure quality mark-up (see Sec.5.2.2), and it should be considered whether this is a task for journalists, editors, or maybe even dedicated mark-up specialists to do.

27 http://iptc.org/std/rNews/2011-10-07#

5.3.2 News article query and user interface

Once all concepts are semantically annotated, it is possible to develop an application to perform complex and complete 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 for their archive of articles – consisting of news articles from 1981 to today – where more than 100.000 people, places, organisations, and descriptors (terms) are annotated (N.Y.T. Developer, n.d.). The application is designed as an Application Programming Interface (API) for public use over the HTTP protocol and is used as a case study in this examination.

N.Y.T.'s semantic API includes seven concept types (see Table 11 below) each represented by a URI-reference in the N.Y.T.-controlled vocabulary[28]. Similarly, the specific concept name – that is searched for – needs to exist as a term in the controlled vocabulary.

Concept type	Concept type name	Specific concept name (example)
Descriptor	nytd_des	Baseball
Location (places)	nytd_geo	Kansas
Organisation	nytd_org	3M Company
Person	nytd_per	Obama, Barack
Public organisation	nytd_porg	Department of Agriculture
Title (creative work)	nytd_ttl	The Joy of Painting (Tv Program)
Topic	nytd_topic	Health

Table 11 N.Y.T. concept types

28 N.Y.T control a range of URI's for their vocabulary. The base URI for one of them looks like this: http://topics.nytimes.com/top/news/

N.Y.T.'s list of concept types almost matches the concept types identified in the qualitative analysis: Persons, organisations, locations, and key terms (see Sec.4.3).
N.Y.T. provides two concept types for organisations: nytd_org and nytd_porg. And three concept types for key terms: nytd_des, nytd_ttl, and nytd_topic. The terms nytd_org and nytd_porg can possibly be mixed up: If a journalist wanted to get an overview of articles about a specific public organisation but used the concept type nytd_org in a search query, theoretically no result would be retrieved. Similarly, it is difficult for users to distinguish between the use of nytd_des and nytd_topic. It is recommended to design a search application with the simplest possible set of concepts (see Table 12 below).

Concept type	Concept type name
Person	media_per
Organisation	media_org
Location	media_loc
Topic	media_topic

Table 12 Recommended set of concept types

Concept types and concept type names can be used to construct a request URI. E.g. if a journalist wants to get an overview of all N.Y.T.'s news articles about Denmark, the semantic request URI should follow this structure:
http://api.nytimes.com/svc/semantic/v2/concept/name/nytd_geo/Denmark&api-key=[API-key][29]

The query result is received in JSON (see Appendix 9.21 for results retrieved

29 To use the N.Y.T. Semantic API, users must register in order to get a personal API key. This allows N.Y.T. to control who uses their service and for what purposes.

on April 2nd, 2020), which is a format commonly used to transmit data between server and Web application.

Fig.18 below shows a comparison between the N.Y.T.'s semantic search result and the result of a similar search on Google which is the search tool journalists most often use for this type of searches (see Sec.4.2).

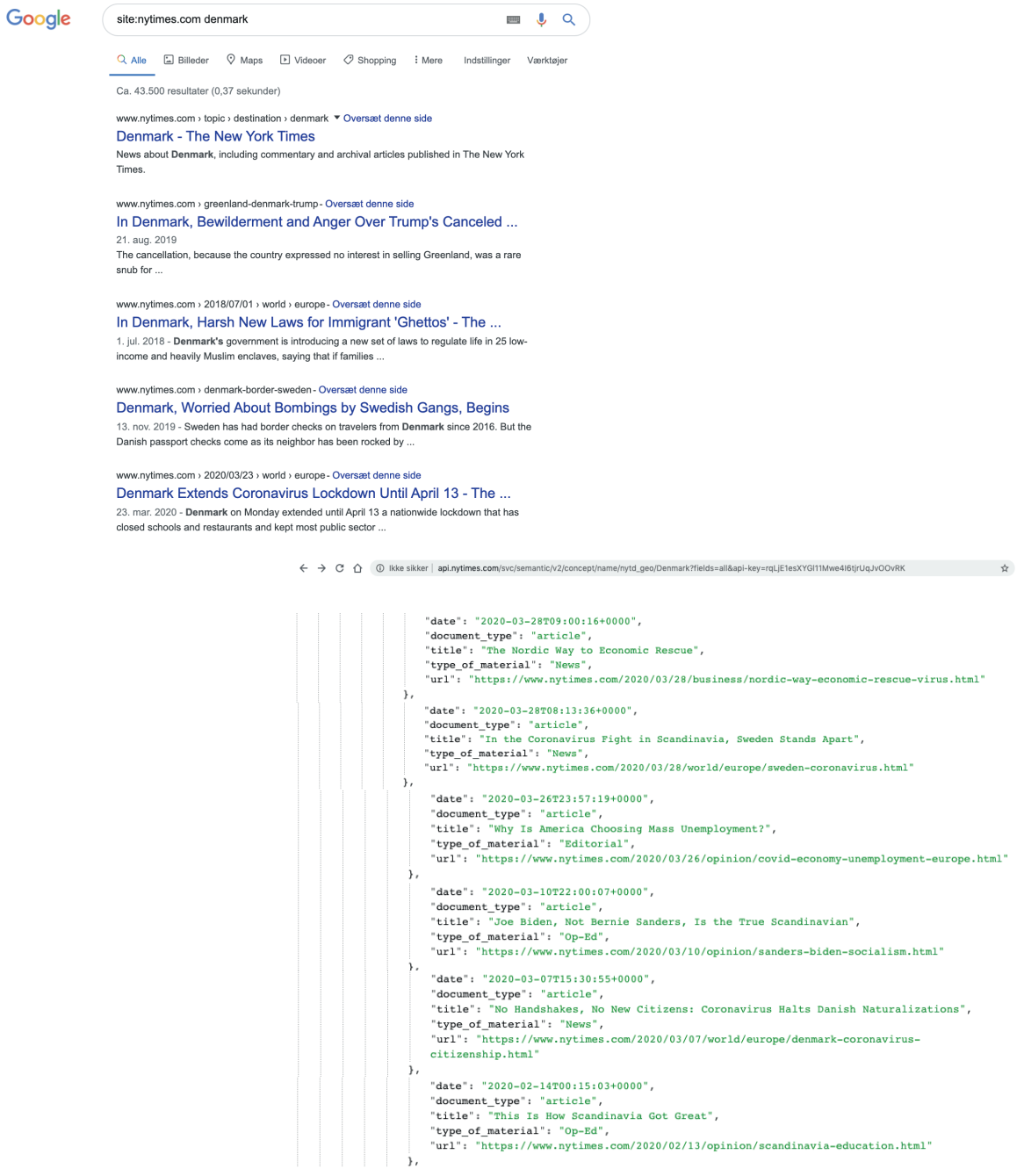


Fig.18 N.Y.T. semantic news article search vs. traditional Google search. The two types of searches are conceptualised in Fig.05

Fig.18 illustrates how N.Y.T.'s semantic search provides a precise and more complete list of articles sorted by the article most lately published, whereas the Google results list is of a more random character.

Journalists cannot be expected to remember the exact URI syntax and spelling of the different concept type names offhandedly. Thus it is recommended to design a search panel and user interface to guide the construction of these queries. Figure 19 below illustrates how an ideal user interface can be designed in order to combine the needs of news journalists and the requirements for a search URI.

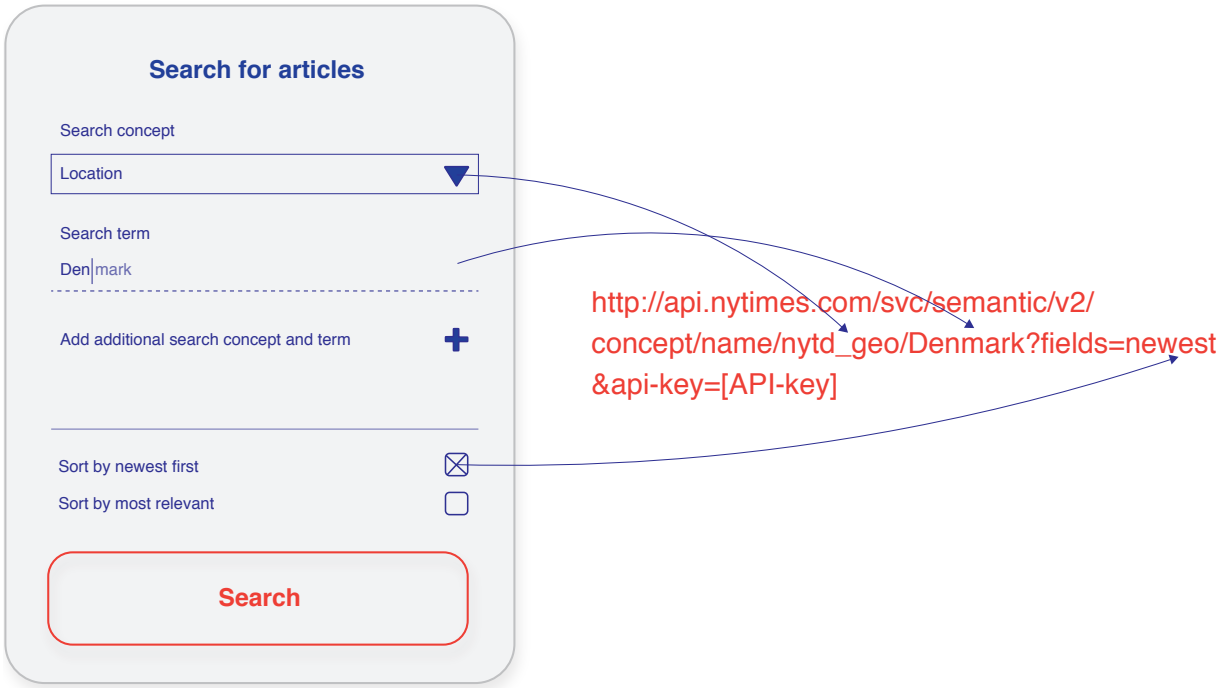


Fig.19 Search panel for semantic news article search

The search panel contains a drop-down menu where users can choose one of four predefined concept types. It is then possible to type a search term and choose between suggested predefined terms (within the chosen search concepts) from the media's controlled vocabulary. These appear from a drop-down list.
Similar to the application for sources and contact details and for the same

reasons, the user interface is phrased in English (see Sec.5.2.3).
Finally, the search panel includes possibilities for combining two or more search concepts and search terms as well as possibilities for sorting the result list by either most recently published articles (newest) or articles most related to the search term (most relevant).
The N.Y.T. Semantic API provides additional optional parameters such as concept_uri and article_uri, however these do not provide information with relevance to news journalists, and the options have thus been left out of a recommended search panel design.

Construction of the N.Y.T. semantic search as an API is sensible as the function is then easy to integrate in different services such as a CMS, an intranet, or even in the front end of a media's public website.
APIs can be implemented in HTML-documents, and results can be retrieved and displayed using RESTful requests (see Sec.3.5) as illustrated in Snippet 08 below.

```
<script>
    var url = "http://api.nytimes.com/svc/semantic/v2/concept/name/nytd_geo/Denmark?fields=newest&api-key=FhAPhIS6jMxs8pvK86fG6OwUiixT4HTq";
...
    $.ajax({
      url: url,
      method: 'GET',
    }).done(function(result) {
            var article = result.response.docs;

            for(i=0; i< 6; i++){
            //checking if the article has an image
            if (article[i].multimedia.length < 3){
            //no image display empty <div>
            var image = "<div></div>";
            }
            else {
            //if there is an image place the image information inside a single variable image
                var image = "<a href='http://www.nytimes.com/"+article[i].multimedia[1].url+"'> <img src='http://www.nytimes.com/"+article[i].multimedia[2].url+"' class='nyt_thumb' alt='"+article[i].headline.main+" image "+article[i].multimedia[2].subtype+"' title='"+article[i].headline.main+" image "+article[i].multimedia[2].subtype+"' height='"+article[i].multimedia[2].height+"' width='"+article[i].multimedia[2].width+"' ></a>"; // thumbnail is being appended and displayed
            }
            var pub_date_day = String(article[i].pub_date.substring(0, 10));
            pub_date_day = pub_date_day.split('-').reverse().join('-')
            $("#nyt").append("<div class='headline'>"+image+"</br>"+article[i].headline.main+"<p>" + pub_date_day + " | " + article[i].byline.original + "</p>" + "<p class='art_text'>" +article[i].snippet+"</p><a href='" + article[i].web_url + "' target = '_blank'>Read more</a></div>"); // headline, date, byline, snippet, and an URL to the original article is being appended and displayed
            }

    }).fail(function(err) {
      throw err;
    });
</script>
```

Snippet 08 Implementation of N.Y.T. semantic search API

This example retrieves the latest N.Y.T. news articles about the descriptor

Olympic Games (l.6 in Snippet 08).

The analysis proves that journalists perform this type of search to get an overview of related articles including headline, description, publication date, and URL-link to the original article (see Sec.4.2). Snippet 08 above demonstrates how this information can be formatted and displayed for each search result using JavaScript (l.21–38 in Snippet 08).

It is possible to display additional information such as related concept types or concept type names, but the analysis does not find a need to include such additional information as journalists do not orientate in topic pages or tags.

Finally, it is recommended to layout search results as a list overview. Fig.20 below demonstrates how three search results – from the query performed in Snippet 08 on April 2nd, 2020 – can be listed.

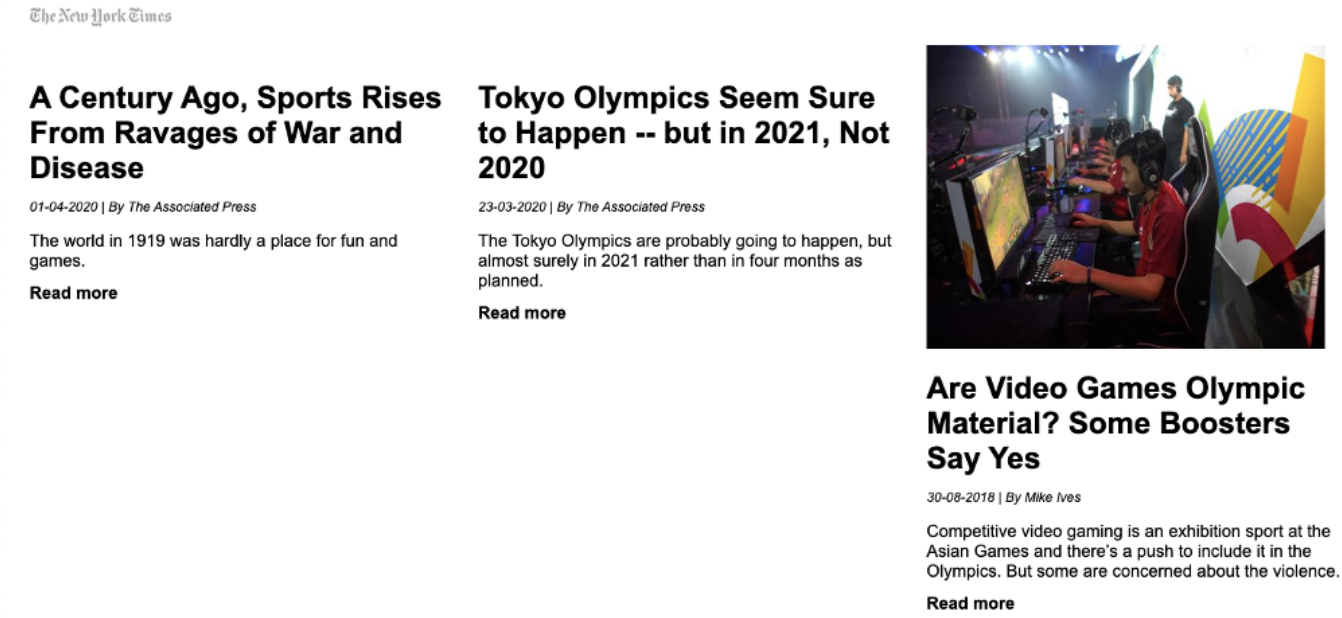


Fig.20 N.Y.T. semantic news article search – recommended display of results

With slight adjustments – as described and discussed in the paragraphs above – it is recommended to develop an internal semantic news article search based on the model of N.Y.T.'s already existing semantic API search.

The section below further discusses the use of a semantic news article search in the context of Danish news media and the cost of semantically annotating every news article.

5.3.3 News article search: Challenges and further development

As described in the sections above, implementation of a semantic news article search application requires semantic annotation of every paragraph in every article published by a Danish news media. This requires training and implementation of completely new work processes for news journalists or even recruitment of annotation specialists. Such implementation is extensive and should be topic for a cost benefit analysis: What are the benefits of a semantic news article search application?
As described in Sec.4.3, an internal semantic search application highly supports the first phases of journalists' research processes. This applies to all of a media's journalists across all domains, and the time spent on searching for relevant news articles might even match the extra time required for annotating key terms and concepts. A complete annotation also strengthens the search tool for sources and contact details (see Sec.5.2) as persons, organisations, and topics will be interweaved in more detail.
Additionally, the search application (in combination with the search tool for sources and contact details) potentially offers completely new possibilities for journalists to search the archive and extract information. Some of these possibilities meet the requirements and wishes expressed by participants in the qualitative research interviews and are worth briefly mentioning.

First of all, the semantic search tool can be used to display relevant articles related to a specific topic. Possibly, this function can automatically insert links to related articles in a specific news article. This is something journalists currently do manually – because existing technologies for inserting related articles are not good enough (see Sec.4.3 and Appendix 9.11). An automation of this process saves time for each journalist and enables links which are not

present to the journalist to be added.

Participants also suggest that a semantic search tool can potentially be used to analyse how minorities are represented and used as sources in the overall news coverage.

> Are ethnic minorities only interviewed for articles about crime statistics? Or are they also used as cases for more ordinary topics like childcare or delays in public transportation? It would be highly relevant for any media to be aware of this type of bias, I think.
> (Participant 01)

Technically, this can be realised by adding a function to the search application that allows users to display what concept type names (resources) most often cooccur in articles describing a specific term, location, person or organisation. As described in the quote above, this type of analysis can be used as a tool for the media, but it can also provide research for new articles worth reading for the public. Participant 02 imagines:

> It would be interesting to analyse a specific topic... as an example, the debate about violation and offending behaviour – it would be interesting to analyse the gender, age, and education of experts used in articles about that. Then maybe we could write something like: "Yong female gender researchers dictate debate about offending behaviour" (...).
> (Participant 02)

It can be concluded that complete annotation of all news articles can potentially benefit all phases of the journalistic work process – from brainstorming on new ideas, to research and publication.
Annotating a media's archive of news articles using URIs also has the advantage of being a lightweight archive. In the qualitative research, Participant 05 mentions that the media he presents has removed options for content search

within the CMS as the entire system then crashes (see Sec.4.3). Such problems are eliminated when applying Linked Data Principles and semantic annotation. As discussed above (see Sec.5.3.2), linked data Web archives can even be managed inside APIs and implemented in different user interfaces.

The sections above describe how a semantic search application can be implemented within a media's own archive of articles. Journalists however often research in other media's archives too (see Sec.4.3), and it would increase usability remarkably, if not only one but all Danish news media organisations decided to semantically annotate their articles and provide open API search tools. If all annotated information in all Danish news media archives live up to the Linked Data Principles (see Sec.3.2), it is possible to interlink the archives and query all of them using the same search panel.

As discussed in Sec.5.2.3, it is recommended to connect resources phrased in Danish to equivalent resources in English as this potentially allows data to be interlinked not only with other Danish news media archives but to be part of one global graph (see Sec.3.3) containing all the world's media archives. Potentially, this allows journalists to query information in news media archives all over the world which might reveal completely new patterns and support journalist cooperation across borders.

5.4 Semantic infobox: Summary

The following sections contain step-by-step examination and discussion of how an application for autogenerated semantic infoboxes can be realised. The examination pivots around development of a summary as a specific type of infobox displaying date of birth, title, work experience, education etc. of a specific person. The aim of this type of application is for journalists to easily

integrate encyclopaedic and trustworthy information about persons mentioned in current news articles.
The section concludes with a discussion on challenges and perspectives on further development.

5.4.1 RDF graphs and vocabularies for semantic summaries

The analysis finds (see Sec.4.4) 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 (see Sec.4.4–4.5).
Additionally, summaries should include personal information such as date of birth, family relations and historical information about previous jobs and memberships. This allows readers of the news article to easily recap information about the person (see Sec.4.4), and ultimately news journalists can use the application as research tool.

The following sections examine how this type of information can be annotated in a media's archive of articles and how its database of linked data can be extracted and used for autogenerating summaries. The use of internal data exclusively 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. Expansion of the application and additional use of external sources are later discussed in Sec.5.4.4.

As described above, a minimum of information is needed for generating infor-

mative summaries. Fig.21 below describes a generic example of this information as an RDF graph and demonstrates how Schema.org can be used to describe most of the data included. The graph can be seen as an extension of the graph displayed in Fig.09 for sources and contact details.

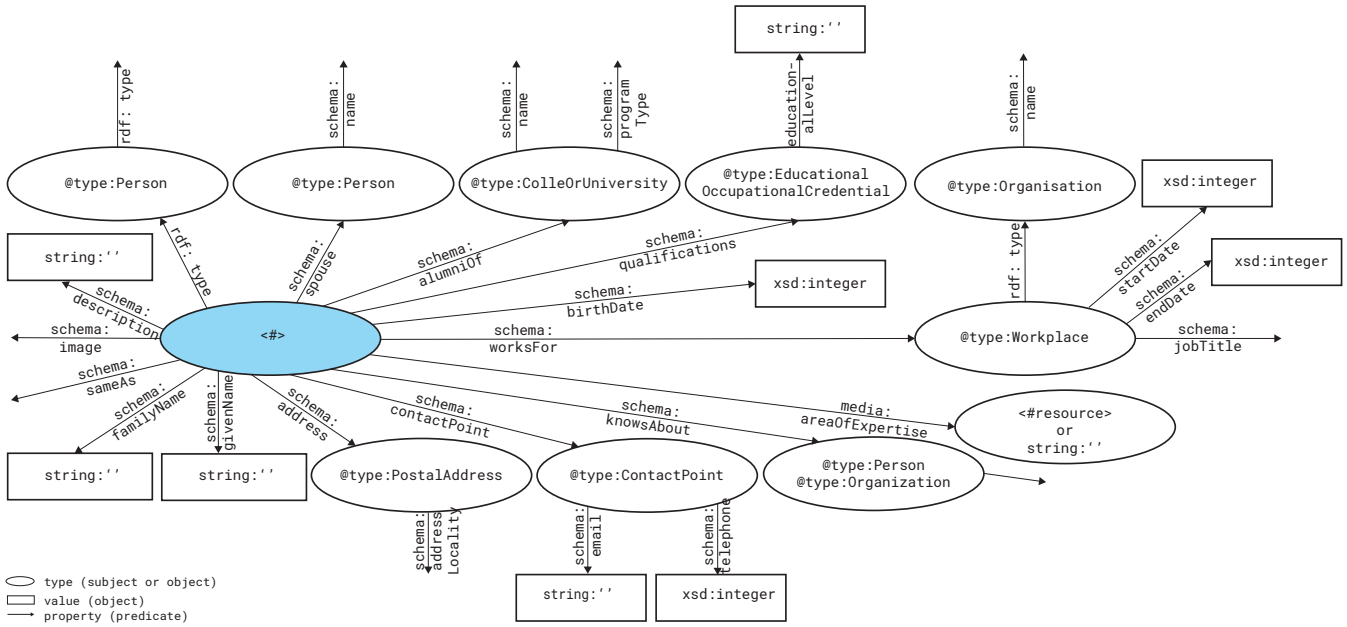


Fig.21 Minimum graph structure for semantic summaries

Schema.org is used in advantage of FOAF (see Sec.5.3.1) as this vocabulary is more detailed when it comes to describing work relations. In the example above, the relation worksFor is applied, this can however be exchanged with relations such as ownerOf or memberOf if the person described is a company owner or member of a political party.

Fig.21 above contains one worksFor-relation. Most people work several places throughout their carrier, and realistically several worksFor-relations are needed. In order to describe employments chronologically, Schema.org recommends[30] using the properties startDate and endDate for each Workplace. Same principle can be applied to the alumniOf-relation to indicate when a Person

30 https://schema.org/Person - see recommendation next to the column hasOccupation

started a specific program or graduated. In this case, startDate- and end-Date-relations must be applied to the programType.
Relations such as spouse, alumniOf and qualifications might be left as blank nodes for some people. These should still be included as part of the summary as no university degree or no spouse is also valuable information.
Finally, the property description contains a string-value as object. This can be used to add and include information written as plain text. Journalists and editors should however aim at including all information as RDF-triples as data can then be reused in other contexts, e.g. to query all Persons related to a specific Workplace or all Persons born on a specific birthDate.

For this type of application, it is important to make sure 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 sameAs-relations (see Sec.5.2.1). 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 (see Fig.23).

All of the information is included in the RDF/XML description of each person (see Sec.5.2.2) forms the data points which the semantic application uses to autogenerate a person's summary.

5.4.2 User interface and fact-checking

An application for autogenerated summaries differs from the two previous solutions, as it contributes with information displayed directly as part of the news article and not only support research. Ultimately, autogenerated content can be implemented by a few clicks, and the journalist should not need to spend additional time on fact-checking and proofreading. 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 (see Sec.4.4).

In 2019, some Danish news media organisations started experimenting with autogenerated content including fully automated articles reporting sport results and financial accountings. In at least one case[31], the media was criticised for publishing crude and deceptive information about local enterprises. The case has been discussed in public media and was unsolicited mentioned by four out of eight participants during the qualitative interviews as something to avoid. These initial experiences with autogenerated content make it clear that journalists are expected – both by users, sources and journalists – to perform source criticism even when they do not write the content themselves. Infoboxes and small summaries differ from traditional news articles, as they can be described as encyclopaedic and as a genre do not declare to present new information. It is however still recommended that journalists are provided with possibilities to check and edit autogenerated summaries before publishing.

Possibilities for fact-checking can be provided as metadata about when the summary was last updated, and what sources the data is collected from. As described in Sec.5.2.2 this kind of transparency and traceability can be implemented using Semantic Sitemaps which allows metadata to be included in the RDF/XML descriptions (see also Sec.3.6).
Fig.22 below demonstrates how a user interface – integrated in the CMS – for implementing and fact-checking autogenerated summaries can be designed. A technical description of how summaries are queried and formatted can be found in Sec.5.4.3.

31 https://www.dr.dk/nyheder/regionale/syd/robotter-spytter-artikler-ud-i-flere-store-medier-hans-otto-endte-i-avisen

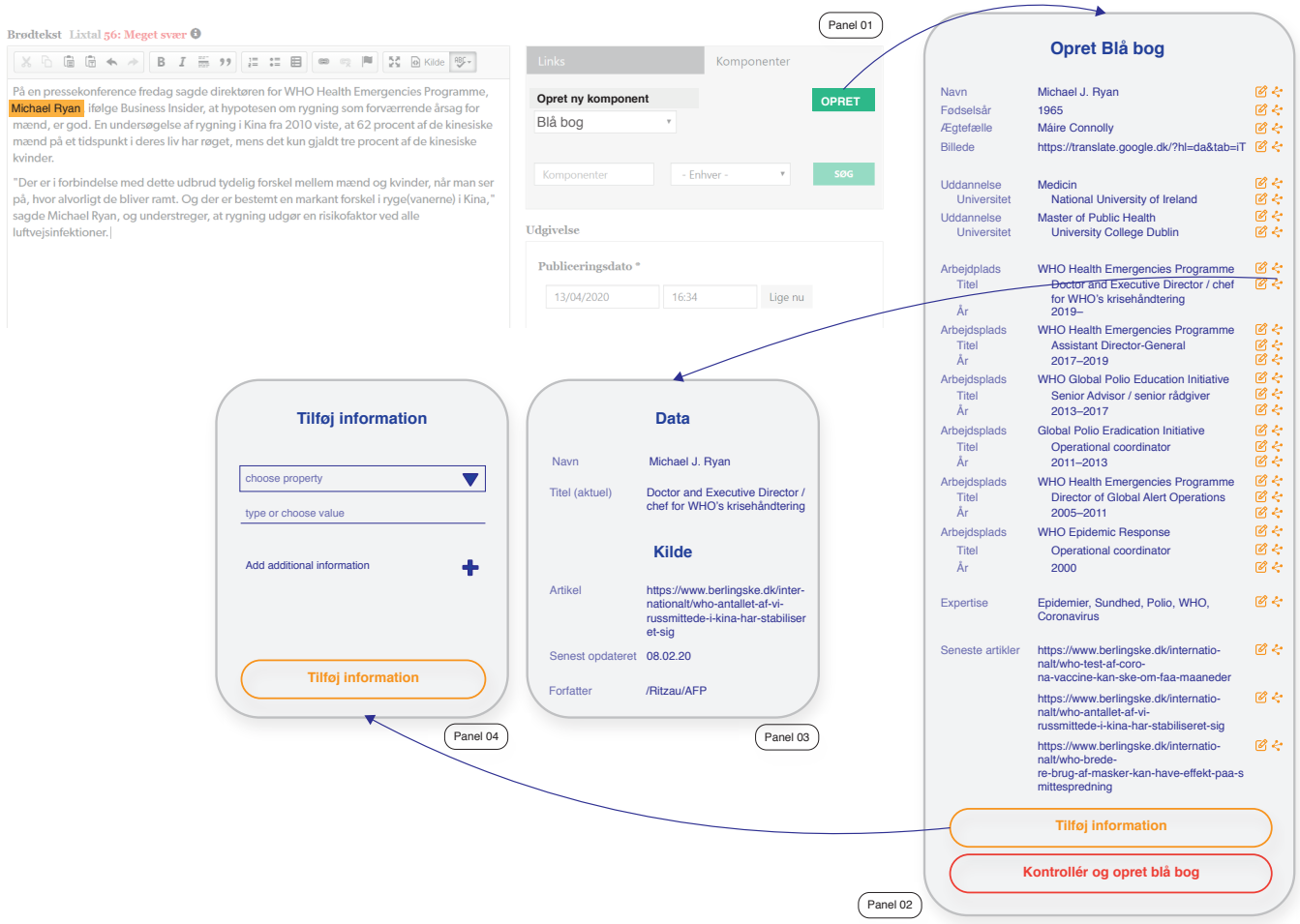


Fig.22 User interface for integrating semantic summaries

The user interface is designed for Danish news media organisations and journalists and the interface is phrased in Danish: E.g. Summary is translated to Blå bog. Similarly, properties and values are preferably retrieved and displayed in Danish, but if the semantic annotation only includes values in English these are being displayed in English. This is often the case for names and job titles.

The analysis finds that journalists are not willing to invest much time on generating summaries, thus a simple, fast-to-use interface is prioritised. The journalist simply highlights a person's name and uses the right-side panel[32]

32 This panel is an existing part of Drupal's user interface and can also be used to paste photos, code boxes, videos etc. into the article.

(panel 01 in Fig.22) to create a summary. A display then appears (panel 02 in Fig.22) 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 (panel 03 in Fig.22). 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 (panel 04 in Fig.22) using predefined properties and predefined or typed values. This panel is phrased in English similar to the user interface in Fig.13 and for similar reasons (see Sec.5.2.3).
When the journalist has reviewed the information, she simply clicks Kontrollér og tilføj blå bog, and an autogenerated summary (see Fig.23) is pasted to the article.

5.4.3 Displaying infoboxes

Summaries can be queried and formatted using SPARQL-queries and SPARQL Lib the same way as results are retrieved and displayed for the sources and contact details application (see Sec.5.2.3). For autogenerated summaries it is recommended to apply a fixed standard query instead of basing it on input from a user panel. Snippet 09 below illustrates how such a standard query for summaries can retrieve information from the media's database of linked data (l.83, Snippet 09) and how it can be formatted. The purpose of this examination is to demonstrate what information should be included and how it can be formatted rather than describing in technical detail how this can be implemented.

```
81          <?php
82            require_once('sparqllib.php');
83            $db = sparql_connect('http://media.dk/sparql');
84            $query = "
85
86            PREFIX dbo:<http://dbpedia.org/ontology/>
87            PREFIX sch:<http://www.schema.org/>
88            PREFIX media:<http://www.media.dk/property/>
89
90            SELECT ?name ?surname ?work ?title ?des ?spouse ?uni ?degree ?topic ?resort ?born.
91            WHERE{
92                ?person a dbo:Person.
93                ?person sch:name Michael_Ryan.
94
95                ?person sch:givenName ?name.
96                ?person sch:familyName ?surname.
97                ?person sch:worksFor ?work.
98                ?person sch:jobTitle ?title.
99                ?person sch:description ?des.
100               ?person sch:spouse ?spouse.
101               ?person sch:alumniOf ?uni.
102               ?person sch:qualifications ?degree.
103               ?person sch:knowsAbout ?topic.
104               ?person media:areaOfExpertise ?resort.
105               ?person sch:birthDate ?born.
106           OPTIONAL {?name ?surname ?work ?title ?des ?spouse ?uni ?degree ?topic ?resort ?born}.
107           FILTER (lang(?name)="da").
108           }
```

Snippet 09 SPARQL request for semantic summary

The example in Snippet 09 queries information about names, title, work organisation, description, spouse, university, degree, area of expertise, and date of birth (l.95–105, Snippet 09) for a Person with the name Michael Ryan (l.92–93, Snippet 09). The OPTIONAL clause (l.106, Snippet 09) makes sure that the summary does not break if some of the information do not exist.
Finally, the query includes a FILTER clause (l. 107 Snippet 09) which filters information to be displayed in Danish, e.g. chef for WHO's krisehåndtering instead of Executive Director (see Fig.22 and Fig.23).

It is relevant to include previous articles – published by the media – related to the person presented in the summary (see Fig.23). This can be achieved using the same method as for querying and displaying articles related to a specific topic (see Sec.5.3.2 and Snippet 08). In this case, topic should be replaced with media_per as concept type (see Table 12), and the name of the specific person as concept type name.
Information retrieved from the SPARQL-query can be formatted and displayed

using PHP and different if-statements as roughly illustrated in Snippet 10 below.

```php
if($result = sparql_query($query)){
  $fields = sparql_field_array($result);
  while($row = sparql_fetch_array($result)){
      echo "<div>";
    foreach($fields as $field)
    {
          echo "<h3>" . $str . "</h3>";
      }
          else
          if ($field == "title"){
          echo "<b>" . $str . "</b>";
      }
          else
          if ($field == "work"){
          echo "<b>" . $str . "</b>";
      }
          else
          if ($field == "des"){
          echo "<b>" . $str . "</b>";
      }
          else
          if ($field == "degree" "uni"){
          echo "<b>" . $str . "</b>";
      }
          else
          if ($field == "title"){
          echo "<b>" . $str . "</b>";
      }
          else
          if ($field == "work"){
          echo "<b>" . $str . "</b>";
      }
        else
          if ($field == "born"){
          echo "<b>" . $str . "</b>";
      }
        else
          if ($field == "spouse"){
          echo "<b>" . $str . "</b>";
      }
          else
          if ($field == "resort"){
          echo "<b>" . $str . "</b>";
      }
          else
          if ($field == "topic"){
          echo "<b>" . $str . "</b>";
      }
          ?>
```

Snippet 10 PHP to integrate and format summary information

Personal summaries already exist as a type of infobox in news articles (see Sec.4.4) and can be considered a genre of its own. When layouting and displaying autogenerated summaries existing standards and conventions should therefore be followed in order to strengthen usability. Fig.23 below illustrates how a summary can be displayed and formatted in a news article context. The example contains the same information as queried and formatted in Snippet 09 and Snippet 10.

På en pressekonference fredag sagde direktøren for WHO Health Emergencies Programme, Michael Ryan, ifølge Business Insider, at hypotesen om rygning som

Michael J. Ryan

Chef for WHO's krisehåndtering

Uddannelse

Medicin, National University of Ireland

Master of Public Health, University College Dublin

Karriere

2019–	Chef for WHO's krisehåndtering WHO Health Emergencies Programme
2017–2019	Assistant Director-General WHO Health Emergencies Programme
2013–2017	Senior rådgiver WHO Global Polio Education Initiative
2011–2013	Operational coordinator Global Polio Eradication Initiative
2005–2011	Director of Global Alert Operations WHO Health Emergencies Programme
2005–2011	Operational coordinator WHO Epidemic Response

Seneste artikler

WHO: Test af coronavaccine kan ske om få måneder

WHO: Antallet af virussmittede i Kina har stabiliseret sig

WHO: Bredere brug af masker kan have effekt på smittespedning

Privat

Født	1965
Partner	Máire Connolly

Ekspertområder

Epidemier · Sundhed · Polio · WHO · Coronavirus

Fig.23 Design and layout of semantic summary

The analysis finds (see Sec.4.4) that users should be provided the opportuni-

ty to either display or skip additional summaries, and it is recommended to integrate summaries as pop-ups[33] when clicking on or simply hovering over the name of a person as it first appears in the news article. Similar to integrated hyperlinks, which are often underlined or highlighted in bold, this requires textual layout, so that users know where to expect additional information. Pop-ups are usually generated using simple HTML, CSS, and JavaScript. Snippet 11 below illustrates how a pop-up can be animated with a simple script-tag to show the pop-up on click.

```
<head>
<style>
    .popup{
        position:relative;
        display:inline-block;
    }
    .popup .popuptext{
        visibility:hidden;
        width: 80%;
        background-color:grey;
        border-radius:6px;
        padding:10px;
        position: absolute;
        z-index: 1;
    }
    .popup .popuptext::after{
        content:"";
        position: absolute;
        top: 100%;
        left: 50%;
    }
    .popup .show{
        visibility: visible;
        -webkit-animation: fadeIn 1s;
        animation: fadeIn 1s;
    }
    @-webkit-keyframes fadeIn{
        from {opacity: 0;}
        to {opacity: 1;}
    }
    @keyframes fadeIn{
        from {opacity: 0;}
        to {opacity: 1;}
    }
</style>
</head>
<body>
    <p>På en pressekonference fredag sagde direktøren for WHO Health Emergencies Programme,
     <span class="popup" onclick="popFunction()">Michael Ryan<span class="popuptext" id="myPopup">
     </span></span>, ifølge Business Insider...
    </p>
    <script>
        function popFuntion() {
            var popup = document.getElementById("Popup");
            popup.classList.toogle("show");
        }
    </script>
</body>
```

Snippet 11 JavaScript to display semantic summary as pop-up

33 Originally, pop-ups are forms of online advertising and consists of small windows that suddenly appear (pops up) in the foreground of the visual interface (Oxford Dictionary, lexico; pop-up).

For clarity, the popup is left without content in the example above (l. 41, Snippet 11). In reality, information retrieved from the SPARQL-query should be inserted inside the second set of span-tags.

5.4.4 Semantic infobox: Challenges and further development

Autogenerated summaries as described in the sections above rely exclusively on linked data from the media's own archive, meaning that the application is limited to display information about persons previously described by the media. Similarly, some people might only be superficially described, which causes the autogenerated summary to be incomplete as datapoints are missing. The information contained is still correct, but the level of inchoate information is a great disadvantage of the application.

In order to truly support news journalists, and to truly take advantage of the concept of Semantic Web, the application should be extended to rely also on external sources of linked data. As an example, access to the open linked database DBpedia (see Sec.3.5) would enable the application to generate summaries (in English) for more than 1.4 million people[34]. The application's trustworthiness is however no longer guaranteed as DBpedia is an open source database that can be edited by anyone.

In other words; external databases must be trusted and protected in order to be included in the application. Furthermore, they must also contain annotated values written in both English and Danish in order to generate and display summaries in Danish.

34 The English version of the DBpedia knowledge base describes 4.58 million things, out of which 1.445.000 are classified as persons (https://wiki.dbpedia.org/about).

Linked databases from public institutions[35] such as the parliament, municipalities, and universities are obvious to include. Even though these institutions are reliable sources, 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 are reviewed at least once every month, and that a person registered with name and contact details is responsible for this maintenance. This might seem like an unproportionally large amount of work to ensure the quality of simple encyclopaedic information; however, it stresses how complex it is to ensure and protect a news media's trustworthiness.
Establishing national or even international standards for the quality and maintenance of linked databases allows media organisations – and other organisations – to share and reuse encyclopaedic data from each other. As demonstrated in the analysis, news journalists across Danish news media organisations work in very similar ways and with extended focus on research and fact-checking (see Fig.08). These conventions might ease the process of defining a set of standards for summary data, but legal and practical implementation of such certifications requires further research.

The sections above, demonstrate how autogenerated summaries can be developed and implemented as a specific type of semantic infobox. A similar method can be applied to develop different kinds of infoboxes to automatically describe organisations, explain key terms, or create timelines. It is out of the scope of this study to describe these additions in detail; but some reflections are worth briefly mentioning:
First of all, the language issue – with annotation values in both English and Danish – is enhanced in any other type of infobox. For the summary, this issue

35 The national platform for public Open Data, OpenData.dk, lists around 900 linked datasets covering topics such as traffic flow counts, locations of public parking meters, drinking water stations, and population prognosis.

is less distinct as English job titles and names of organisations and educations are often used in the Danish language. It would however be highly distracting if an infobox provides information about a term or an invention in a mix of Danish and English.
The context of information is also difficult to incorporate in autogenerated content. The same way as news articles present the most relevant information first, infoboxes are often edited to support the news article in the best possible way. This editing is difficult to implement automatically, and the information provided might seem less relevant.
Finally, decisions about layout and design should be considered. The summary is lay outed as a simple table, though more complicated designs are required to communicate timelines.

5.5 Partial conclusion (answering RQ3)

Chapter 5 examines and discusses what technical requirements and usability considerations can be found for three Semantic Web applications to be realised:

- Application I: Semantic archive of sources and contact details
- Application II: Internal semantic news article search
- Application III: Semantic infobox: Summary

Each application includes unique features; however, the applications also share several common traits and requirements:

Each application requires semantic annotation of a large number of concepts (persons, places, organisations, and key terms) described in a media's archive of articles. It is recommended that standard vocabularies are used to write this annotation as URIs, and underlying rules and taxonomies can then be reused. The vocabulary Schema.org has proved to be applicable for describing persons, their personal details, relations to other persons and to organisations. In addition to Schema.org domain specific vocabularies – such as

http://purl.org/NET/c4dm/event.owl# **and** http://www.w3.org/2003/01/geo/owl# **– needs to be integrated to describe events, sport activities, geographical locations etc. Additionally, the examination identifies one case where standard vocabularies need to be extended with a new property (**areaOfExpertise**) in order to describe the relation between a source and this person's area of expertise. For the semantic article search, technical analysis demonstrates that annotation of metadata is beneficial, and that this can be applied using the IPTC's rNews ontology.**

By default, most standard vocabularies phrase resources and properties in English. This however demonstrates to be inconvenient in the context of Danish news media where query results and autogenerated content should be displayed in Danish in order to ensure usability. Technical analysis indicates that this inconvenience can be managed by creating URIs phrased in Danish – within the media's controlled namespace – for each resource and property. Ideally, each of these Danish resources **and properties should be connected to English equivalences (in standard vocabularies) via** sameAs**-relations to ensure that the annotation becomes part of the global graph constituting the Semantic Web.**

Annotated linked data can be retrieved using SPARQL-queries or API technologies in combination with RESTful requests.
Queries can either be implemented as fixed pre-set SPARQL-queries – as is the case for the semantic summary **– or be constructed based on input from a user interface – as is the case for** the archive of sources and contact details **as well as** the semantic news article search**. Journalists are not trained in writing SPARQL-queries and the importance of a robust user interface to guide the construction of each search query is emphasised. The user interface must cater for both technical requirements such as SPARQL syntax and request URIs as well as the news journalists' possible needs. PHP and different** if-statements **can be used to format and display retrieved information.**

In theory, it is possible to annotate, and query linked data for the three types of Semantic Web applications, but severe challenges are met when it comes to apply semantic annotation to every paragraph in every news article. It is recommended to write the annotation as RDFa immediately before publishing, and to support this process by semi-automatic software. This type of software – e.g. Calais or N.Y.T.s Editor – exists today, but further research is needed to examine efficiency and accuracy in a Danish context.

Technical analysis demonstrates that two of the three applications proposed in this study – the Internal semantic news article search and the Semantic infobox: Summary – require profound annotation of all persons, places, organisations, and key terms mentioned in a media's archive of news articles. Even with support from annotation software, this process is highly time consuming and requires implementation of completely new work processes, contributions from several news journalists, and possibly employment of dedicated annotation specialists.
The third application – the Semantic archive of sources and contact details – requires less thorough annotation and is thereby a more realistic suggestion for a Semantic Web application that can benefit Danish news media in near future.

The technical discussion concludes that it is beneficial to apply metadata about the semantic annotation for all application types to ensure traceability and reliability of data. Metadata can be provided as Semantic Sitemaps and should include information about when the annotation was last updated and who wrote it. Potentially, this metadata can reinforce a shared set of standards or a type of certification to guarantee trustworthy linked data, which ultimately enables Danish news media – and other organisations – to share and reuse annotated information.

Finally, it is important to stress that the three applications proposed in this

study are developed to support – not replace – journalists. They should be considered tools similar to encyclopaedias or calculators, and journalists using the application should always be critical towards the information provided and the context in which it is to be used.

6.0 Comparison and outlook

Chapter 6 discusses key findings of this study in perspective of the current state of journalism– and Semantic Web research. This is done to evaluate this study's conclusions and to qualify its contribution to the fields of research. Sec.6.1 compares the Semantic Web applications proposed in the study at hand with existing solutions developed for news media, while Sec.6.2 discusses methods for future research and development. Finally, Sec.6.3 discusses aspects of AI.

6.1 Comparison with existing Semantic Web applications for news media

As part of the introductory literature review of this study (see Sec.1.1.1), 25 existing Semantic Web applications for the news media industry has been identified (see Appendix 9.2).
A majority of these applications (see Table 13 below) are concerned with content search and new ways of presenting already published information. As an example, BBC's music page enables users to list all BBC programmes featuring a specific artist.
Only five of the applications are developed as research tools to support journalists' work process.

Type of Semantic Web application	Number of identified examples
Application for categorising, sorting, or presenting published content	13
Application allowing users to generate new content by querying information across multiple datasets	6
Application designed as research tool for journalists	5 (two of these are not developed as prototypes but merely exist as theoretical thought-up examples)

Application designed to commercially distribute and sell journalistic information	1

Table 13 Categorisation of Semantic Web applications for the news media industry. Categories have been applied using a grounded theory approach.

This predominantly user-centred focus can be interpreted as a result of a developing process where insights of journalists and editors have not previously been implemented. Instead the needs of end-users (consumers of news stories) have been put in focus.

In comparison, all of the applications proposed in this study (see Sec.5.2–5.4) can be characterised as research tools that in some way support the work process of Danish news journalists.

A previous focus on developing Semantic Web applications for categorising and sorting published information should also be understood in the context of when these applications were developed. Most of the applications identified in the literature review are launched around 2007 (see Appendix 9.2). At this time standard search engines were facing problems with ambiguous search terms and the integration of multiple search results (see Sec.1.1.2). Since then, the PageRank-algorithm has become much more powerful, and most lately Google Search also applies structured data to enable special search result features (Google, 2020). As a result of this rapid development, a large part of the issues described to motivate the development of Semantic Web applications for news media (see Sec.1.1.2) are no longer applicable as Google already bridge that gap.

This comparison indicates, that it might be beneficial to change the objective of Semantic Web development for the news media industry: Instead of focusing on how already published information can be presented, this study de-

monstrates (see Sec.4.2–4.5) that potentials of Semantic Web are more likely to be unfolded within the work process of news journalists.

This perspective also contributes to journalism research. In most newsrooms, ways of digital storytelling are being discussed and examined (Meadows, 2003). A majority of these discussions concern superficial styling and how content can be animated and communicated interactively (Pavlik & Pavlik, 2017). However, Semantic Web technologies – as presented in this study – introduce not only ways of animating and communicating information, but new practices for content creation, including models for how journalists can use previous research and articles as mouldable building blocks instead of a completely blank sheet every time news stories break.
This constitutes entirely new areas of journalism research and potentially leads to separate examinations of the future role of journalists, cooperation between media organisations, documentation, and reliability.

6.2 Domain-oriented development

This study applies a domain-oriented approach for the development of Semantic Web applications. First, needs within a specific domain – Danish news media – have been analysed, then Semantic Web standards, methods and techniques have been combined and applied to accede those needs.
The approach is acknowledged within IT development for its ability to clarify circumstances, requirements, and difficulties specific to a certain domain (Oliveira, Rocha, & Travassos, 1999). Within the field of Semantic Web the approach is however relatively new. Instead, Semantic Web research tends to focus on developing universal standards and on describing how these in theory can be applied in all contexts (see Sec.3.1–3.4).

In this study, the approach fruitfully illuminates how the guarantee of trustworthy information should be paid special attention when developing applications for Danish news media, and that this cannot yet be fully achieved within the concept of Semantic Web (see Sec.5.4.4).
Another major domain-specific challenge identified in this study is the practical annotation of large amounts of text material (see Sec.5.2.2). In this case, existing theories are applicable, but robust software still needs to be developed before it can be achieved in a time efficient way.

The findings of this study are not extensive enough to conclude anything substantial, but they indicate that it might be fruitful to apply a domain-oriented approach in future Semantic Web research. When examining and developing for one domain at the time, researchers are forced to solve challenges of a more pragmatic character which according to this study is what is needed for the Semantic Web to truly make a leap.

6.3 Trust and Proof – and AI

Artificial Intelligence (AI) can be described as

> (...) phrases that are intimately associated with the very essence of the Semantic Web vision
> (Idehen, 2001).

The core intention of Semantic Web is to make content readable by machines which ultimately not only improves findability but also enables 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 very distant. As described in Sec.5.4.2, remarkably little research

explores aspects of reliability and objectivity within the field of Semantic Web even though major challenges on how to secure trustworthy information need to be solved before AI can even be discussed. As demonstrated in Sec.5.4.4 these challenges need to be solved not only locally, but as standards or certifications agreed upon by all trustworthy linked data providers. Before even suggesting implementation of AI agents, Semantic Web researchers must provide solutions for these top layers of the Semantic Web Stack.

The concept of neutral ontologies (see Sec.3.4) is also worth briefly mentioning in the context of AI. The concept of ontologies presupposes that the entire world can be objectively categorised. However, several current debates illustrate how this in reality often becomes a matter of political discussion, e.g. should the term Homosexuality be classified as a subclass of the class Illness or not? The World Health Organisation defines homosexuality as an illness, while Danish health authorities do not.

This illustrates how 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 traceable and should be an area of extreme interest for researchers in the field of Semantic Web and in the field of journalism.

7.0 Conclusion

Chapter 7 concludes this study and summarises its findings and discussion points. The chapter concludes with remarks on limitations and recommendations for future work in Sec.7.1.

This study is addressed to examine how the work process of Danish news journalists and the user experience of Danish news journalism can be improved in the context of Semantic Web?

News media has long constituted an area of interest for Semantic Web researchers, but remarkable little research combines the knowledge of technologists with insights of editors and journalists. To fill the gap in literature, this study applies a qualitative research approach – including qualitative interviews and PD studies with eight journalists and editors – to clarify in what areas Semantic Web technologies can potentially contribute to Danish news journalism anno 2020.

The examination reveals three areas with significant potential of improvement. The first area concerns journalists' challenge of finding the right person to comment on or evaluate a specific topic. The second area concerns issues of finding previously published articles related to a specific concept. Finally, the third area targets the need for generating additional, encyclopaedic infoboxes in a short amount of time.

Each of these areas is analysed in the context of Semantic Web and has been translated into three types of Semantic Web applications:

- **Application I: Semantic archive of sources and contact details**
- **Application II: Internal semantic news article search**
- **Application III: Semantic infobox: Summary**

Each application requires semantic annotation of a large number of concepts (persons, places, organisations, and key terms) mentioned in a media's archive of news articles. It is recommended to use standard vocabularies to apply this annotation as URIs and underlying rules and taxonomies can then be reused.

By default, most standard vocabularies phrase resources and properties in English, however, this demonstrates to be inconvenient in the context of Danish news media where query results and autogenerated content should be displayed in Danish in order to ensure usability. Technical analysis indicates that this inconvenience can be managed by creating URIs phrased in Danish – within the media's controlled namespace – for each resource and property. Ideally, each of these Danish resources and properties should be connected to English equivalences (in standard vocabularies) via sameAs-relations to ensure that the annotation becomes part of the global graph constituting the Semantic Web.

Once annotated, linked data can be retrieved using SPARQL-queries possibly in combination with API technologies.

In theory, it is possible to annotate, and query linked data for the three types of Semantic Web applications proposed in this study. One however faces severe challenges when it comes to applying semantic annotation to large bodies of text. It is recommended to write the annotation as RDFa immediately before publishing, and to support this process by semi-automatic software. This type of software – such as Calais or N.Y.T.s Editor – exists today, but further research is needed to examine efficiency and accuracy in a Danish context.

Technical analysis demonstrates that two of the three applications proposed in this study – the Internal semantic news article search and the Semantic infobox: Summary – require profound annotation of all persons, places, organisations, and key terms mentioned in a media's archive of news articles. Even

with support from annotation software, this process is highly time consuming and requires implementation of completely new work processes.
The third application – the Semantic archive of sources and contact details – requires less thorough annotation and is thereby a more realistic suggestion for a Semantic Web application that can benefit Danish news media in near future.

Another major challenge is the issue of securing trustworthiness and proof of documentation. Reliable information is crucial for any news media but is currently an Achilles' heel for the concept of Semantic Web.
Technical discussion on these topics concludes that it is beneficial to apply metadata for the semantic annotation for all application types to ensure traceability. Metadata can be annotated as Semantic Sitemaps and should as minimum include information about the author and when the annotation was last updated. Idealy, this type of metadata can reinforce a type of certification guaranteeing trustworthy linked data, which ultimately enables Danish news media – and other organisations – to share and reuse annotated information.

The study achieves to answer the overall research question and can be considered as groundwork for future research. The examination describes in detail in which areas Danish news media can potentially benefit from Semantic Web technologies and demonstrates how robust user interfaces and existing technologies can be applied to develop three types of Semantic Web applications.

Even though Semantic Web has been considered open for business (Miller, 2008) for more than a decade, this study demonstrates that several areas still require basic research before theory can be turned into practice and benefit Danish news media. This especially includes the practical implementation of annotation software as well as aspects of trust and proof of documentation (see recommendations for future research in Sec.7.1).

As a final remark, it is recommended to examine RQ2 – how can the user experience of Danish news journalism be improved by Semantic Web technologies? **– in more detail. In this study journalists and editors are used as informants, but their representation of news media users is poor, and this part of the examination would benefit from interviews and PD studies with a more diverse group of participants.**

7.1 Concluding remarks on limitations and recommendations for future work

The aim of this study is to bridge technical knowledge and domain-specific insights from news journalists. This scope contributes with novice approaches to the body of Semantic Web research, but it also contains limitations.

The study can be considered a type of case study, meaning that the areas of potential improvement identified in this examination is not an exhaustive exploration. There might be several other areas and processes within Danish news journalism where Semantic Web technologies can contribute. These can be examined in detail by zooming in on an even narrower aspect of Danish news journalism, such as reporting on politics or crime, or by zooming out and collecting empirical material from more respondents across the industry.
The scope of this study also limits detailed discussions on how different solutions can be implemented and executed. This is a disadvantage as it can be concluded that practical aspects of Semantic Web technologies are currently insufficiently described and need to be paid more attention by researchers in the field.

More specifically, this study demonstrates that basic research on how annotation can be applied in a Danish context is needed. Especially challenges of combining

annotation in different languages seem to be neglected in the body of research. Finally, it can be concluded that the top layers of the Semantic Web Stack – including layers of Unifying Logic, Proof, and Trust – calls for fundamental research. This becomes especially obvious when Semantic Web applications are discussed in the context of news media: Before news media organisations can truly take advantage of Semantic Web and external RDF-links, global standards on how to guarantee and protect trustworthy information is needed. This includes traceability of metadata as well as research on how transparency in the construction of ontologies can be achieved.

8.0 References

Adida, B., Birbeck, M., McCarron, S., & Herman, I. (2015). Syntax and processing rules for embedding RDF through attributes. Retrieved from https://www.w3.org/TR/rdfa-core/

Alam, F., Rahman, S. U., Khusro, S., & Ali, S. (2015). Towards a Semantic Web Stack Applicable for Both RDF and Topic Maps: A Survey. Technical Journal, 20(II).

BBC Newslab. (2018). **The Juicer**. Retrieved from https://bbcnewslabs.co.uk/projects/juicer/

Berners-Lee, T. (2005). Uniform Resource Identifier (URI): Generic Syntax. Retrieved from https://tools.ietf.org/html/rfc3986

Berners-Lee, T. (2006). Linked Data. Retrieved from https://www.w3.org/DesignIssues/LinkedData.html

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web – A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. Scientific American Magazine, 29–37.

Blaikie, N. (2007). Approaches to Social Enquiry - Advancing Knowledge (2nd ed.). Polity Press.

Blois, M., Escobar, M., & Choren, R. (2007). Using agents and ontologies for application development on the semantic web. Journal of the Brazilian Computer Society, 13(2).

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 (pp. 145–182). Routledge.

Brandt, E., & Grunnet, C. (2000). Evoking the future: Drama and props in user centered design. Proceedings of the Conference on Participatory Design, 1, 11–20.

Bryman, A. (1988). Quantity and Quality in Social Research. Unwin Hyman Ltd.

Buchenau, M., & Suri, J. F. (2000). Experience Prototyping. Proceedings of the Conference on Designing Interactive Systems: Processes, Practices, Methods and Techniques, 424–433.

Bødker, K., Kensing, F., & Simonsen, J. (2004). Participatory IT Design: Designing for Business an Workplace Realities. Cambridge, MA: MIT Press.

Creamer, M. (2008). It's Web 3.0, and someone else's content is king. Advertising Age, 79(15).

Domingue, J., Fensel, D., & Hendler, J. (Eds.). (2011). Handbook of Semantic Web Technologies (Vol. 1). Berlin: Springer.

Feitosa, D., Dermeval, D., Farias Lóscio, B., & Isotani, S. (2017). A systematic review on the use of best practices for publishing linked data. Online Information Review. Retrieved from https://doi.org/10.1108/OIR-11-2016-0322

Finlayson, A. (2010). The Peril and Promise of the Semantic Web. Nieman Reports.

Goddard, Lisa & Byrne, G. (2010). Linked Data tools: Semantic Web for the masses. First Monday, 15(11). Retrieved from https://firstmonday.org/ojs/index.php/fm/article/view/3120/2633Slide

Google. (2020). Understand how structured data works. Retrieved from https://developers.google.com/search/docs/guides/intro-structured-data

Gubrium, J. F., & Holstein, James, A. (2003). Postmodern Trends in Interviewing. In Postmodern Interviewing. SAGE Publications.

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.

Idehen, U. K. (2001). A Semantic Web & Artificial Intelligence. Scientific American Magazine.

Jaques, Y., Anibaldi, S., Celli, F., Subirats, I., Stellato, A., & Keizer, J. (2012). Proof and Trust in the OpenAGRIS Implementation. Int'l Conf. on Dublin Core and Metadata Applications.

Jarhi, A. Al, & Gaaly, T. (2007). Trust and Proof in the Semantic Web. Retrieved from http://www.cse.aucegypt.edu/~csci585/StudentsProjectsSpring07/Jarhi&GaalyReport.pdf

Kidder, L. H., & Judd, C. M. (1986). Research Methods in Social Relations. Holt, Rinehart and Winston.

Kobilarov, G., Scott, T., Raimond, Y., Oliver, S., Sizemore, C., Smethurst, M., Lee, R. (2009). Media Meets Semantic Web – How the BBC Uses DBpedia and Linked Data to Make Connections. Retrieved from https://doi.org/10.1007/978-3-642-02121-3_53

Koivunen, M.-R., & Miller, E. (2001). W3C Semantic Web Activity. Retrieved from https://www.w3.org/2001/12/semweb-fin/w3csw

Kyrnin, J. (2020). Why Use Semantic HTML? Retrieved from https://www.lifewire.com/why-use-semantic-html-3468271

Meadows, D. (2003). Digital Storytelling: Research-based practice in new media. SAGE Publication, 2(2).

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

Oliveira, K. M., Rocha, A. R., & Travassos, G. H. (1999). A Domain-Oriented Software Development Environment for Cardiology. AMIA Annual Symposium.

Pandey, R., & Sanjay, D. (2010). Interoperability between Semantic Web Layers: A Communicating Agent Approach. International Journal of Computer Applications, 12(3).

Pavlik, J. V., & Pavlik, J. O. (2017). Understanding Quality in Digital Storytelling: A Theoretically Based Analysis of the Interactive Documentary. Digital Transformation in Journalism and News Media, 381–396.

Poole, D., Mackworth, A., & Goebel, R. (1998). Computational Intelligence and Knowledge. In Computational Intelligence: A Logical Approach. New York: Oxford University Press.

Raimond, Y., Scott, T., Oliver, S., Sinclair, P., & Smethurst, M. (n.d.). Use of Se-

mantic Web technologies on the BBC Web Sites. Retrieved from https://doi.org/10.1007/978-1-4419-7665-9_13

Ritchie, J., Lewis, J., Nicholls, C. M., & Ormston, R. (2003). Qualitative Research Methods: A Data Collector's Field Guide Qualitative Research Methods Overview. SAGE Publications.

Robson, C. (2002). Real World Research. Blackwell Publishing.

Rogers, Y., Preece, J., & Sharp, H. (2015). Interaction Design. Beyond Human -Computer Interaction. Chichester: John Wiley.

Rowley, J. (2007). The wisdom hierarchy: Representation of the DIKW hierarchy. Journal of Information and Communication Science, 33(2), 163–180.

Sandhaus, E. (2012). Administering a 160-year-old Database Knowledge Management at New York Times. Retrieved from http://videolectures.net/solomon_sandhaus_administering/

Silverman, D. (2011). Interpreting Qualitative Data. A Guide to the Principles of Qualitative Research. SAGE Publications.

Sizov, S. (2007). What Makes You Think That? The Semantic Web's Proof Layer. IEEE Computer Society.

Spradley, J. (1979). The Ethnographic Interview. Holt, Rinehart and Winston.

Tanggaard, L., & Brinkmann, S. (2010). Interviewet: Samtalen som forskningsmetode. In Kvalitative metoder: En grundbog (pp. 29–53). Hans Reitzels Forlag.

Uschold, M., & Gruninger, M. (2004). Ontologies and semantics for seamless connectivity. SIGMOD, 33(4), 58–64.

Veglis, A., & Bratsas, C. (2017). Reporters in the age of data journalism. Journal of Applied Journalism & Media Studies, 6(2).

Wood, D., Marsha, Z., Luke, R., & Hausenblas, M. (2013). Linked Data - Structured Data on the Web. Manning Publications.

Yin, R. K. (1994). Case study research: Design and Methods (2. Ed.). SAGE Publications.