Semantic Metadata in the News Production Process – Achievements and Challenges

Tassilo Pellegrini
St. Pölten University of Applied Sciences
Matthias Corvinus Str. 15
A-3100 St. Pölten, Austria
tassilo.pellegrini@fhstp.ac.at

ABSTRACT

Only a few media companies have so far (as of April 2012) publicly declared engagement in the area of Linked Data. Nevertheless among the chosen few are BBC Online, the New York Times, The Guardian and Reuters who utilize Linked Data to add significant value to the news production process. This paper discusses achievements and challenges in utilizing semantic metadata in the news production process. By discussing a BBC use case it illustrates how Linked Data can be integrated into the content value chain and provide added value to content-related workflows without necessarily disrupting them. This is insofar critical as publishing companies react very sensitively to radical changes in their working settings and are often very suspicious of technologically induced innovations. Given the fact that from the perspective of media professionals Linked Data is a highly technology driven phenomenon that incrementally incorporates the culture, speech and logic of the engineering discipline, it is necessary to translate the benefits of Linked Data into the thinking and understanding of the publishing sector by illustrating the intersections between the traditional editorial content value chain and Linked Data as a complementary resource to innovate existing products and services.

Categories and Subject Descriptors

E.0 [**General**]; K.4.3 [**Organizational Impacts**]; K.5 [**Legal Aspects of Computing**]

General Terms

Management, Economics, Human Factors, Standardization, Legal Aspects.

Keywords

Linked Data, Content Value Chain, Semantic Metadata, Semantic Web, Data Journalism, News Production, Editorial Workflows, Media Economics, IPR, Data Licensing

MindTrek 2012, October 3-5, 2012, Tampere, FINLAND.

1. Introduction

The increasing availability of highly structured data as part of open data strategies by governments, companies or crowd-sourcing initiatives drives the question of how the content and publishing industry can benefit from this trend and incrementally integrate this new kind of resource into their production and business processes. Although we have seen conceptual work about the value chain related to the production of Linked Data [33], no attention has so far been paid to the question how Linked Data contributes to existing value chains in the content industry. This paper tries to close this gap by pointing to existing achievements and discussing a BBC (British Broadcasting Corporation) use case in the utilization of semantic metadata for the management of news content along the content value chain.

Experience shows that due to risk aversion, lack of financial resources and expertise actors in the media industry tend to behave very cautiously when it comes to the adoption of new technologies and methodologies of content creation and reuse, especially when they carry a strong disruptive potential and affect their core business, competencies or corporate culture. But with the renaissance of data driven approaches towards journalistic content creation – i.e. Data Journalism [1; 6] – the acceptance towards automated data analytics and visualization methods which support the time critical creation of new knowledge, is rising.

Given the fact that media professionals perceive Linked Data as a highly technology driven phenomenon [52] that incrementally incorporates the culture, speech and logic of the engineering discipline it is necessary to translate the benefits of Linked Data into the thinking and understanding of the media sector by illustrating the interfaces between the traditional editorial content value chain and Linked Data as a complementary resource that adds value to existing products and services. For reasons of validity this will be exemplified by focusing on the news sector as this branch of the media industry is especially sensitive to technological innovations and its disruptive potential to existing business models and editorial production cultures. This is insofar ironic as the news industry has been among the early adopters of the World Wide Web for publishing purposes but very few actors have succeeded in adjusting to the “open and collaborative spirit of the Web” as exercised by technology-driven market entrants like Google, Yahoo! or Facebook that are now being perceived as the biggest challengers to traditional journalism and advertising based business models.

The paper is structured as follows: Section 2 illustrates the role of metadata in the news production process and highlights the added value when the design principles of Linked Data are being applied to them. Section 3 introduces the concept of the Content Value Chain and gives an overview over academic research done at the intersection of news production and semantic metadata. Section 4 discusses licensing issues in connection with semantic metadata. Section 5 discusses a use case provided by the BBC who use semantic metadata to improve their editorial workflow. Section 6 gives a conclusion and outlook on future research.

2. The Renaissance of Metadata and the News Industry

With the emergence of the World Wide Web as a universal platform for the creation and distribution of information the nature and importance of metadata has changed significantly. To illustrate this Saumure & Shiri [41] conducted a survey on research topics in the Library and Information Sciences by analyzing publications from the LISTA Database (Library, Information Science, and Technology Abstracts)[1] with the year 1993 as a demarcation line of a pre-web and a post-web era. Table 1 shows their research results.

Table 1. Research Areas in Library and Information Science

Research Area	Pre-Web	Post-Web
Metadata Applications / Uses	--	16 %
Cataloging/ Classification	14 %	15 %
Classifying Web Information	--	14 %
Interoperability	--	13 %
Machine Assisted Knowledge Organization	14 %	12 %
Education	7 %	7 %
Digital Preservation/ Libraries	--	7 %
Thesauri Initiatives	7 %	5 %
Indexing/ Abstracting	29 %	4 %
Organizing Corporate or Business Information	--	4 %
Librarians as Knowledge Organizers of the Web	--	2 %
Cognitive Models	29 %	1 %

The survey illustrates three trends: 1) the spectrum of research areas has broadened significantly; 2) while certain areas have kept their status over the years (i.e. Cataloging & Classification or Machine Assisted Knowledge Organization), new areas of research have entered the discipline (i.e. Metadata Applications & Uses, Classifying Web Information, Interoperability Issues) and others have declined or dissolved into other areas; and 3) metadata issues have significantly increased in importance in terms of the quantity of papers that is explicitly and implicitly dealing with corresponding issues.

2.1 Metadata News Standards – Traditions and Pitfalls

The "Metadata Shift" [18] described in the above mentioned survey can also be observed in news-related metadata initiatives like the International Press Telecommunications Council (IPTC)[2] that predominantly releases media-relevant metadata and schema recommendations for areas like news, sports, events, photos and more. According to their self-understanding the IPTC does not only provide "news exchange formats to the news industry but also creates and maintains sets of concepts to be assigned as metadata values to news objects like text, photographs, graphics, audio- and video files and streams, [which] allows for a consistent coding of news metadata over the course of time."

Theses metadata resources are called IPTC NewsCodes[3] and can be distinguished by descriptive codes, administrative codes, transmission codes and exchange-format codes. The corresponding controlled vocabularies (schemata) and taxonomies are based on XML and incorporate so called Globally Unique Identifiers (GUID) that are unique, unambiguous and persistent. Over the recent years some individually created RDF/SKOS-representations of IPTC codes emerged that in the meanwhile are also officially provided and maintained by IPTC. Nevertheless the design and purpose of the IPTC codes is primarily targeted at inhouse use with limited semantic interoperability in terms of http-based de-referenceable URIs. Additionally interviews with industry insiders [52] reveal that the practical uptake of the IPTC codes among the news industry and its usage in editorial content management systems and applications is limited to a small fraction of the existing vocabulary which is a strong indicator for over-specification on the one side and a lack of an elaborated "metadata culture" in the management of information within editorial workflows on the other. There have even been proposals from the scientific community to extend the expressivity of IPTC codes [i.e. 48; 29] but they have not been taken up by industry.

Nevertheless the problems of over-specification and de-referenceability have recently (as of November 2011) been addressed by the IPTC with the release of a semantic web-enabled microformat called rNews v1.0[4]. rNews can be serialized as RDFa or HTML5 Microdata. Its data model is composed of a reasonably manageable amount of concept classes with a sufficient level of semantic expressivity in terms of vocabulary and data types that cover the most important attributes of a news item. If necessary the vocabulary can be extended with other IPTC sources or any other controlled vocabulary that adhere to the RDF specifications. In this respect rNews represents the first serious step towards the implementation of Linked Data principles into institutionally governed metadata infrastructure for the news industry that does not just meet the formal requirements of a machine-readable, semantically interoperable standard but also provides it in a way that meets the requirements of newsrooms.

Beside the IPTC news-relevant metadata sources are also provided by the Newspaper Association of America[5] (i.e. ANPA-

[1] See: http://www.ebscohost.com/academic/library-information-science-technology-abstracts-lista, as of April 4, 2012

[2] See: http://www.iptc.org/site/Home/, visited April 4, 2012

[3] The IPTC NewsCodes consist of the following subsets: EventsML-G2, NewsML-G2, SportsML-G2, rNews, IIM, NewsML 1, IPTC 7901, NITF

[4] See: http://dev.iptc.org/rNews, visited April 4, 2012

[5] See: http://www.naa.org/, visited April 4,2012

1312, a 7-bit news agency text markup) or the microformat hNews[6] as provided by Associated Press. Additional standards[7] commonly used in the media industry are EXIF, Dublin Core, XMP, DIG35, P/Meta, ETSI/TV Anytime or MPEG-7. Nevertheless, most of them are designed for multimedia purposes and play a minor role in the news production process.

2.2 Benefits of Linked (Meta)Data

Semantic interoperability is crucial in building cost efficient IT systems that integrate numerous data sources. Since 2009 the Linked Data paradigm has emerged as a light weight approach to improve data portability among various systems. By building on Semantic Web standards and principles the Linked Data approach offers significant benefits compared to conventional data integration approaches. These are according to Auer [2]:

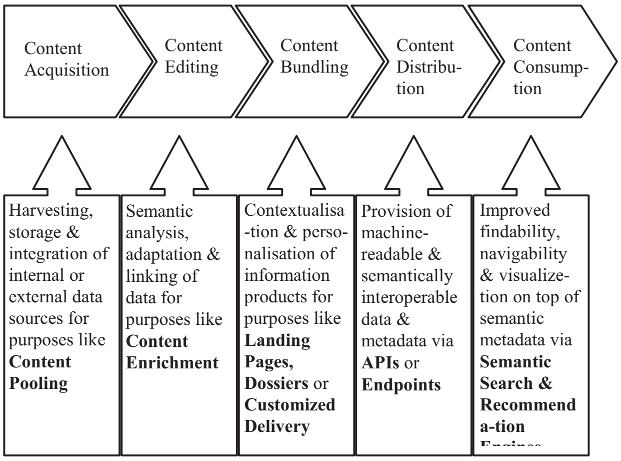
- **De-referencability.** IRIs are not just used for identifying entities, but since they can be used in the same way as URLs they also enable locating and retrieving resources describing and representing these entities on the Web.
- **Coherence.** When an RDF triple contains IRIs from different namespaces in subject and object position, this triple basically establishes a link between the entity identified by the subject (and described in the source dataset using namespace A) with the entity identified by the object (described in the target dataset using namespace B). Through these typed RDF links, data items are effectively interlinked.
- **Integrability.** Since all Linked Data sources share the RDF data model, which is based on a single mechanism for representing information, it is very easy to attain a syntactic and simple semantic integration of different Linked Data sets. A higher-level semantic integration can be achieved by employing schema and instance matching techniques and expressing found matches again as alignments of RDF vocabularies and ontologies in terms of additional triple facts.
- **Timeliness.** Publishing and updating Linked Data is relatively simple thus facilitating a timely availability. In addition, once a Linked Data source is updated it is straightforward to access and use the updated data source, since time consuming and error prune extraction, transformation and loading is not required.

On top of these technological principles Linked Data promises to improve the reusability and richness (in terms of depth and broadness) of news content thus adding significant value to the news production process. Along this line of argument the following section will elaborate how these characteristics can be utilized to add value to the content value chain in the news production process.

3. Linked Data in the Content Value Chain – Related Work

The concept of the value chain is rooted in industry economics, where it is used to provide an analytical framework for business management processes. It has been first introduced by Michael Porter [39] and since then been adopted in a variety of ways to describe the structure of sector specific value creation mechanisms and sequential production logics. According to this notion the news production process can be conceptualized as a value chain generally consisting of five steps: 1) content acquisition, 2) content editing, 3) content bundling, 4) content distribution and 5) content consumption. As illustrated in Figure 1 Linked Data can contribute to each step by supporting the associated intrinsic production function.

Figure 2. Linked Data in the Content Value Chain


Content Acquisition
Content Editing
Content Bundling
Content Distribu-tion
Content Consump-tion
Harvesting, storage & integration of internal or external data sources for purposes like **Content Pooling**
Semantic analysis, adaptation & linking of data for purposes like **Content Enrichment**
Contextualisa-tion & perso-nalisation of information products for purposes like **Landing Pages, Dossiers** or **Customized Delivery**
Provision of machine-readable & semantically interoperable data & metadata via **APIs** or **Endpoints**
Improved findability, navigability & visualize-tion on top of semantic metadata via **Semantic Search & Recommend a-tion**

Beside elaborating on these conceptual thoughts the following sections will present related scientific work along the various steps of the value chain.

3.1 Content Acquisition

Content acquisition is mainly concerned with the collection, storage and integration of relevant information necessary to produce a news item. In the course of this process information and facts are being pooled from internal or external sources for further processing.

At the time of writing there is a limited amount of work on data pooling in connection with journalistic practices and Semantic Web / Linked Data sources. While the underlying "immature integration capabilities" have already been subject to criticism back in 2006 [16] modest approaches exist that address the topic of structured data acquisition from a general perspective [i.e. 21; 17] and mainly in connection with multimedia. Schandl et al. [42] provide a general overview over the possibilities to retrieve media-relevant information from Linked Data sources like DBpedia[8] or the Linked Movie Data Base[9]. Messina et al. [36] describe a Hyper Media News System that makes use of large scale automated data acquisition from RSS feeds for Electronic Programme Guides. Hausenblas [19] and Kobilarov et al. [28] provide an insight how the BBC is pulling Linked Data to

[6] See: http://microformats.org/wiki/hnews, visited April 4, 2012

[7] See: EXIF - http://www.exif.org/; Dublin Core - http://dublincore.org/; DIG 35 - http://www.i3a.org/technologies/digitalimaging/metadata/; P/Meta - http://tech.ebu.ch/metadata/p_meta; TSI/TV Anytime - http://www.etsi.org/website/technologies/tvanytime.aspx; MPEG-7 - http://www.multimedia-metadata.info/, visited April 20, 2012

[8] See: http://dbpedia.org, visited April 10, 2012

[9] See: http://www.linkedmdb.org/, visited April 10, 2012. See also Consens [7]

improve existing web applications for purposes like content syndication, enrichment and page navigation.

Anyway, all of these works refer to the technological feasibility of Linked Data integration but none of them addresses aspects of data quality like provenance, reliability or trustworthiness. But data quality is a crucial factor in the news production process and should be targeted carefully when integration issues are concerned – especially in the context of journalistic practices.

3.2 Content Editing

The editing process entails all necessary steps that deal with the semantic adaptation, interlinking and enrichment of data. Adaptation can be understood as a process in which acquired data is provided in a way that it can be used in the editorial process. Interlinking and enrichment are often performed via processes like tagging and/or referencing to enrich media documents either by disambiguating existing concepts or by providing background knowledge for deeper insights.

Especially the enrichment of media documents in the editorial workflow has been subject to numerous works and is probably the most elaborated area in the utilization of semantic metadata. Early work [37; 30; 45] provides design principles for a metadata life cycle management and demonstrates the value of well-structured metadata for indexing and compiling multimedia documents across various modalities like text, speech and video. Hee et al. [20] illustrate how RSS feeds from various content providers can be used to enrich TV Anytime data and utilize it for news packaging services thus aggregating content from various sources into one stream. Hu et al. [23] present an ontology-based reasoning system that supports the automated analysis and aggregation of news items from very large corpora. Their system is capable of extracting names, terms, temporals (i.e. time stamps and durations) and locations incorporating metadata from OpenCyc, EventsML-G2, NewsML-G2 and News Industry Text Format (NITF) and reason over these elements to detect semantic similarities between news items. Kim et al. [26] discuss the benefits of domain ontologies as integration layers for tagging purposes in collaborative environments like a newsroom. They demonstrate that shared knowledge models can help to improve the reuse of tags from various systems and increase flexibility in the editorial process without disrupting it. Yu et al. [49] describe a system that uses a two-staged SOA-based approach for semantically annotating and brokering TV-related services in Electronic Program Guides. By utilizing the light-weight semantics available from Linked Data they were able to significantly lower the integration complexity of external data and improve the maintenance of the system for developers. And Mannens et al. [35] propose a system for automated metadata enrichment and indexing to improve customized delivery of news items based on the IPTC NewsML-G2 standard. They enrich the IPTC news model with extracted entities from DBpedia and generate facets that help the user to fuzzily browse a knowledge base thus improving content discovery.

The work discussed in this section is of high importance in the editorial news process. Content editing is a highly time- and cost-intensive activity and any measure that helps to reduce the resources needed to produce high quality content are at the core of the news production process. Unfortunately the work discussed above builds on the premise that a sufficient amount of quality approved metadata is already available. As this is rarely the case for news content, further research is necessary on the (semi-)automatic generation of high quality metadata on the one side and its reuse for the semantic adaptation, interlinking and enrichment of news content on the other.

3.3 Content Bundling

The bundling process is mainly concerned with the contextualisation and personalisation of information products. It can be used to provide customized access to media files i.e. by using metadata for the device-sensitive delivery of news items, or to compile thematically relevant material into Landing Pages or Dossiers thus improving the navigability, findability and reuse of information.

In the area of content bundling we find a large amount of work that is dealing with the role of semantic metadata in the development of new products and delivery services. Back in 2001 Jokela et al. [25] have proposed a system called “SmartPush” that adapts to the personal preferences of the user by linking user feedback to the well structured set of metadata used to annotate news items. They applied a Mixed-Initiative-Approach which has outperformed purely algorithmic recommendation services but suffers from a cold-start problem. A similar approach has been proposed by Bomhart [3], Zhou et al. [51] and Gao et al. [13] who utilize behavioral data (i.e. browsing history) to match personal reading preferences with structural similarities between news items. And Knauf et al. [27] present a system that uses a rich set of metadata derived from manual and automated video annotations to discover semantic links and similarities between multimedia files and use this for personalized delivery purposes especially on mobile devices.

Utilizing behavioral data to personalize services, enrich the reading experience and improve advertising relevance has a long tradition in human computer interaction [53; 54; 55]. Anyway, excessive collection and reuse of behavioral data poses threats to user privacy and can cause adverse effects among the targeted users. Additionally these systems often suffer from a cold-start problem that can negatively impact the user experience.

In contrast to these mixed-initiative approaches that match news items to user profiles Liu et al. [34], Bouras & Tsogkas [5] and Scouten et al. [43] illustrate how the extraction of metadata from media files can be used to calculate similarities between media files and thus improve the relevance of automated news selection and recommendation services. They achieve this by combining rule-based natural language processing techniques with domain ontologies thus providing fuzzy-search mechanisms to the user that lets him conveniently explore a news database. This approach is also followed by Ijntema et al. [24] who propose ATHENA, an extension to the HERMES framework [10], used to build news personalization services by combining domain ontologies with text analytics from GATE and WordNet. This work is extended by Goosen et al. [15] who use a Concept Frequency – Inverse Document Frequency algorithm for personalized news services on top of the ATHENA framework.

What the latter approaches have in common is that they utilize domain ontologies to organize metadata and pull reasonable information from linked vocabularies. This allows new forms of knowledge discovery and delivery services that go beyond the established search and retrieval paradigms and provide the users with a richer interaction experience without necessarily intruding their privacy.

3.4 Content Distribution

In a Linked Data environment the process of content distribution mainly deals with the provision of machine-readable and semantically interoperable (meta)data via Application Programming Interfaces (APIs) or SPARQL Endpoints [50]. These can be designed either to serve internal purposes so that data can be reused within controlled environments (i.e. within or between newsrooms) or for external purposes so that data can be shared between unknown users (i.e. as open SPARQL Endpoints on the Web).

From a technical perspective any work dealing with Linked Data Publishing is relevant to this specific step in the content value chain and lots of media-related datasets are already available as Linked Data (i.e. LinkedMDB [7], music and movie related data on DBpedia[10], MusicBrainz[11] or the Linked Media Framework by [31]). Over the past years we have also seen several examples where media companies have started to offer news-relevant data as Linked Data. Since 2009 BBC is offering public SPARQL Endpoints for their program, music and sports data [28; 44; 40] and in the same year the New York Times has started to offer large amounts of subjects headings as SKOS via their Article Search API[12] [32]. The Reuters Open Calais API[13] supports SOAP, REST and HTTP requests providing data in RDF [8]. And The Guardian is offering an MP Data SPARQL Editor[14] from which data about British politicians can be retrieved in RDF [9].

Despite these encouraging examples news companies have not yet embraced the logic and culture of open publishing – especially when data assets are concerned. Here we face a typical chicken-and-egg-problem where the willingness of opening up data sources heavily depends on the quantifiable added value that can be derived from such a strategy. Due to the lack of differentiation possibilities competing news-products are usually substitutes of each other. Thus there is little incentive for news companies to diversify their business models into this direction as they 1) need to protect their revenue channels and 2) often face licensing constraints when reusing content outside of established practices.

3.5 Content Consumption

The last step in the content value chain is dealing with content consumption. This entails any means that enable a human user to search for and interact with content items in a pleasant und purposeful way. So according to this view this level mainly deals with end user applications that make use of Linked Data to provide access to news items i.e. by providing reasonable visualizations.

While most of the methodologies for contextualisation and personalization and corresponding applications like search and recommendation services described in Section 3.2.3 (Content Bundling) serve the purpose of content consumption, increasing attention has also been paid to visualization and interaction issues associated with Linked Data although most of it is still in an experimental phase. A comprehensive overview is given by Paulheim [38], who himself introduced a tool called Semantic Data Explorer. And Freitas et al. [11] discuss recent challenges, approaches and trends in querying and visualizing heterogeneous datasets on the Linked Data web. Anyway, tools that specifically target the needs of journalists or that make explicit use of news-related content are still very rare.

Fu et al. [12] provide conceptual thoughts how to visualize time series data from consumer-generated media and news archives but do not explicitly touch upon the benefits of Linked Data. Böhm et al. [4] present a tool for journalists called GovWild that integrates and cleanses open government data at a large scale thus providing easy access to linked government data via a simple search interface. A similar approach is followed by Hoxha et al. [22] who provide a research tool for journalists to visualize Albanian Government Data using the Google Visualization API to process RDF data retrieved via an easy to use SPARQL endpoint.

Research on and the improvement of interface design for the handling of semantic data services will be one of the crucial success factors in the broad adaptation of Linked Data in the news environment. Tools and functionalities need to be designed to meet the requirements and IT skills of technology lay persons. Technology should be harboured from them as good as possible.

4. Beyond Standards & Technology – Licensing Linked Data

New technology has never been a sufficient precondition for the transformation of business practices but has always been accompanied by complementary modes of cultural change [14]. In the case of Linked Data one of these non-technological modes is the availability of alternative legal instruments (i.e. commons-based approaches) that allow the owner of intellectual property rights to design a licensing environment that suites the attributes and specificities of the digital asset concerned and allows to establish new modes of value creation.

4.1 Traditional IPR Instruments

Semantic metadata is a fairly new kind of intellectual asset that is still subject to debate concerning the adequate protection instruments [47]. Table 2 gives an overview over the applicability of various IPR instruments. The table illustrates the complex nature of semantic metadata as intellectual property. Various instruments can be applied to various assets, while copyright, database right and competition right are the most relevant ones.

Copyright basically protects the creative and original nature of a literary work and gives its holder the exclusive legal right to reproduce, publish, sell, or distribute the matter and form of the work. Hence, any literary work that can claim a sufficient degree of originality can be protected by copyright.

Database Right protects a collection of independent works, data or other materials which are arranged in a systematic or methodological way and are individually accessible by electronic or other means. Databases are also protected as literary works and need to have a sufficient degree of originality that requires a substantial amount of investment.

An **Unfair Practices Act** protects rights holders against certain trade practices, which are considered unfair in terms of misappropriation, advertising, sales pricing or damages to reputation. Especially the first aspect is relevant to semantic metadata, which actually occurs, when data is being reused without appropriate compensation i.e. in terms of attribution or financial return.

[10] See http://dbpedia.org/About, visited April 20, 2012

[11] See http://musicbrainz.org/, visited April 20, 2012

[12] See http://developer.nytimes.com/docs, visited April 20, 2012

[13] See http://www.opencalais.com/documentation/calais-web-service-api, visited April 20, 2012

[14] See http://api.talis.com/stores/guardian, visited April 20, 2012

Patenting does not directly impact the protection of semantic metadata as – at least in Europe – patents can just be acquired for hardware-related inventions. But as soon as semantic metadata becomes indispensable subject of a methodology that generates physical effects, has a sufficient level of inventiveness and can be exploited commercially these components can be protected under Patent Law.

Table 2. IPR Instruments for Semantic Metadata

	Copyright	**Database Right**	**Unfair Practice**	**Patents**
Documents	YES	YES	YES	NO
Base Data	NO	NO	PARTLY	NO
Description	YES	NO	YES	NO
Identifier	NO	YES	NO	NO
Name Space	YES	YES	YES	NO
Vocabulary	PARTLY	YES	YES	NO
Classification	PARTLY	PARTLY	PARTLY	NO
Ontology	PARTLY	YES	YES	PARTLY
Rules	PARTLY	YES	YES	PARTLY

4.2 Commons-based Approaches

The open and non-proprietary nature of Linked Data design principles allow to easily share and reuse this data for collaborative purposes. This also offers new opportunities to news publishers to diversify their assets and nurture new forms of value creation (i.e. by extending the production environment to open or closed collaborative settings) or unlock new revenue channels (i.e. by establishing highly customizable data syndication services on top of fine granular accounting services based on SPARQL).

To meet these requirements commons-based licensing approaches like Creative Commons[15] or Open Data Commons[16] have gained popularity over the last years, allowing maximum re-usability while at the same time providing a framework for protection against unfair usage practices and rights infringements. Nevertheless to meet the requirements of the various asset types, a Linked Data licensing strategy should make a deliberate distinction between the database and the content stored in it. This is necessary as content and databases are distinct subjects of protection in intellectual property law and therefore require different treatment and protection instruments. An appropriate commons-based protection strategy for a data provider could look as follows:

- The **contents** of a linked dataset, which are comprised of the terms, definitions and its ontological structure, are protected by a CC-By v3.0 License[17], which allows the commercial and non-commercial reuse of any published artefact as long as the owner is mentioned.
- The underlying **database**, which is comprised of all independent elements and works that are arranged in a systematic or methodological way and are accessible by electronic or other means, are protected by a ODC-By v1.0 License[18], which also allows the commercial and non-commercial reuse of any published artefact as long as the owner is mentioned.
- Additionally to these two aspects the licensing strategy also should incorporate a **Linking Policy Community Norm**, which explicitly defines the expectations of the rights holder towards good conduct when links are made to the various artefacts provided in the dataset.[19] This norm should provide administrative information (i.e. creator, publisher, license and rights), structural information about the dataset (i.e. version number, quantity of attributes, types of relations) and recommendations for interlinking (i.e. preferred vocabulary to secure semantic consistency).

All in all the three elements of a commons-based licensing policy - the CC-By v3.0 License, the ODC-By v1.0 License and the Community Norm - provide a secure and resilient judicial framework to protect against the unfair appropriation of open datasets.

5. BBC's Dynamic Semantic Publishing – A Use Case

The BBC (British Broadcasting Corporation) was among the first media companies to utilize Semantic Web standards and Linked Data principles within their content management system. In 2009 they released the BBC Music Beta website [28] where they exercised semantic content enrichment by pulling large amounts of music related data from open repositories like the MusicBrainz database and DBpedia (the semantically linked database version of Wikipedia). In 2010 they extended this approach by publicly launching the BBC World Cup 2010 website [56] which semantically processed all information related to the soccer world cup 2010, hereby coining the term Dynamic Semantic Publishing. This approach has been further extended to the BBC's Olympic Games 2012 website [57] to improve the editorial process for journalists and provide the audience with a semantically powered rich media experience. According to the value chain approach described above their solution works as follows[20]:

- **Content Acquisition:** Editorial content is pulled from over 700 internal sites that contain sports related information and a couple of automated XML sports feeds from various external sources that provide statistics on matches, teams and players.
- **Content Editing:** The editing process is powered by a sports ontology that contains all concepts and relations relevant to the domain. The ontology's concepts are being de-referenced against linked data sources like GeoNames[21] and DBpedia to disambiguate concepts and ensure semantic soundness. This interlinked ontology is utilized for various purposes. It supports

[15] See http://creativecommons.org/, visited April 22, 2012

[16] See http://opendatacommons.org/, visited April 22, 2012

[17] See http://creativecommons.org/licenses/by/3.0/, visited April 20, 2012

[18] See http://opendatacommons.org/category/odc-by/, visited April 20, 2012

[19] See for example the community norm provided by the Leibniz Information Center for Economics: http://zbw.eu/stw/versions/8.08/mapping/gnd/, visited April 20, 2012

[20] [56 & 57] provide a detailed description of the technical architecture.

[21] See http://www.geonames.org/, visited May 5, 2012

the (semi-)automatic annotation of media files by matching labels from a text against labels in the ontology and suggesting it to the editor for annotation. Additionally after the editor has chosen relevant concepts automated semantic enrichment is exercised by using forward-chaining reasoning which adds hierarchically related concepts to the media file.

- **Content Bundling:** An additional meta-model has been created that represents the editorial structure of the websites. This ontology pulls editorial content (profiles, stories and videos) from the repositories, enriches it with dynamic content (stats for matches, teams and players) and automatically compiles a contextually sound, near-real-time (1 min delay) rich media site for the reader.
- **Content Distribution:** For sharing and syndication purposes the BBC has developed a clear IRI schema through which various assets like pages, indices or stories can be identified adding significantly to the de-referenceable granularity of the BBC content. Additionally it provides its well documented sports ontology for download[22] so that it can be reused for third party purposes. An open API for the retrieval of instance data and multi-platform syndication has not been provided by the time of writing.
- **Content Consumption:** The dynamic bundling of content has been optimized to serve millions of page requests every day providing the audience with multimodal reading experience. The system offers thematically clustered landing pages and provides a deep granularity of content by compiling various information types like text, videos and stats to profiles of athletes, matches and disciplines that can be accessed via a shallow facetted search. A bookmarking functionality allows to compile personal dossiers out of page favourites.
- **Licensing Strategy:** The BBC follows a hybrid licensing strategy. It licenses static content under traditional copyright law but provides its ontologies and metadata under a creative commons CC-BY-3.0 attribution to the public. A Community-Norm for the repurposing of ontologies and metadata is missing.

BBC has announced to extend this technological approach to additional sites and domains. According to their experience "[r]eplacing a static publishing mechanism with a dynamic request-by-request solution that uses a scalable metadata/data layer will remove the barriers to creativity for BBC journalists, designers and product managers, allowing them to make the very best use of the BBC's content." [57]

6. Conclusion & Outlook

Without claiming exhaustivity this paper tried to illustrate the large amount of scientific research accomplished in the utilization of semantic metadata for news-related purposes. Additionally it discussed the issue of Linked Data licensing as a complementary practice to derive economic value from semantic metadata and briefly discussed a use case that illustrates how semantic metadata can be applied within the news production process. But a wider look into newsrooms reveals that there is a mismatch between the scientific progress and the uptake of semantic metadata in the realm of the news industry. One of the reasons lies in the underdeveloped metadata culture within news companies where due to missing competence, dense delivery dates and a lack of convenient service-supported workflows the collaborative maintenance of a well-curated metadata base is simply out of scope. Most of the work discussed in this paper builds on technological preconditions – like available domain ontologies, well annotated media files or competence in natural language processing – that have not yet been integrated into the content and workflow management systems available in newsrooms. Examples like BBC provide valuable inspiration and proof of concepts that semantic metadata can add significant value to the news production process. Hence we might see a gradual shift in the media industries from static to dynamic publishing of content. But as long as Linked Data is not seamlessly integrated into the technological production environments of newsrooms and the efforts of creating and maintaining domain ontologies are not significantly lowered there is little chance that his resource will soon be part of a journalist's daily routine.

Another aspect that has not been touched upon in this paper is the quality assurance of Linked Data with respect to validity, trustworthiness and provenance. Given the fact that journalists are highly sensitive to such quality criteria, efforts have to be undertaken to significantly improve the quality of Linked Data and provide reliable measures for quality maintenance – especially when it is created via crowdsourcing or similar collaborative practices. Tackling this challenge is equally important as managing the technological feasibility of Linked Data in editorial processes but probably much harder to accomplish.

A third aspect that should be mentioned here affects the issue of business and revenue models. Linked Data will make a big leap forward if it can prove to either significantly reduce the costs of news production in existing editorial workflows or create new revenue channels by either adding value to advertising clients or providing incentives for readers to pay for content or services. If and only if one of these aspects can be proven privately owned media will be willing to invest substantial amounts of money into a new generation of network technologies.

Beside any obstacles to realization this paper tried to illustrate that there exists a significant value for the news production process that can be derived from Linked Data. But it still requires a modest leap to bring its benefits from scientific research into industrial practice.

7. REFERENCES

[1] Arana, Ana (2012). Data Now. In: Index on Censorship, 41/1, 2012, p. 178-179

[2] Auer, Sören (2011). Creating Knowledge Out of Interlinked Data. In: Proceedings of WIMS'11, May 25-27, 2011, p. 1-8

[3] Bomhardt, Christian (2004). NewsRec, a SVM-driven Personal Recommendation System for News Websites. In: IEEE/WIC/ACM International Conference on Web Intelligence, 20-24 Sept. 2004, p. 545-548

[4] Böhm, Christoph; Naumann, Felix; Freitag, Markus (2010). Linking open government data: what journalists wish they had known. In: Proceedings of the 6th International Conference on Semantic Systems, ACM, p. 1-4

[5] Bouras, Christos; Tsogkas, Vassilis (2009). Personalization Mechanism for Delivering News Articles on the User's

[22] See http://www.bbc.co.uk/ontologies/sport/2011-02-17.shtml, visited May 5, 2012.

Desktop. In: Fourth International Conference on Internet and Web Applications and Services, 24-28 May 2009, p.157-162

[6] Chambert, Lucy; Gray, Jonathan (2012). The Data Journalism Handbook. New York: O'Reilly Media

[7] Consens, Mariano P. (2008). Managing Linked Data on the Web: the LinkedMDB showcase. In: Proceedings of Latin American Web Conference 2008, 28-30 Oct. 2008, p. 1-2

[8] Cryans, Jean-Daniel; Ratte, Sylvie; Champagne, Roger (2010). Adaptation of Apriori to MapReduce to Build a Warehouse of Relations Between Named Entities Across the Web. In: 2nd International Conference on Advances in Databases Knowledge and Data Applications (DBKDA), 11-16 April 2010, p. 185-189

[9] Dodds, Leigh; Davis, Ian (2009). MP Data SPARQL Editor. In: http://www.guardian.co.uk/open-platform/apps-mp-data-sparql-editor, visited April 20, 2012

[10] Frasincar, F., Borsje, J., Levering, L. (2009). A Semantic Web-Based Approach for Building Personalized News Services. International Journal of E-Business Research, 5/3, 2009, p. 35–53

[11] Freitas, André; Curry, Edward; Oliveira, João Gabriel; O'Riain, Seán (2012). Querying Heterogeneous Datasets on the Linked Data Web. Challenges, Approaches, and Trends. In: IEEE Internet Computing, 16/1, 2012, p. 24-33

[12] Fu, Tak-chung; Sze, Donahue C.M.; Leung, Patrick K.C.; Hung, Kei-yuen; Chung, Fu-lai (2007). Analysis and Visualization of Time Series Data from Consumer-Generated Media and News Archives. In: Proceedings of Web Intelligence and Intelligent Agent Technology Workshops (WI-IAT), 2007 IEEE/WIC/ACM, p. 259-262

[13] Gao, Feng; Yuhong Li; Li Han; Jian Ma (2009). InfoSlim: An Ontology-Content Based Personalized Mobile News Recommendation System. In: 5th International Conference on Wireless Communications, Networking and Mobile Computing, 24-26 Sept. 2009, p.1-4

[14] Ghosch, Rishab Ayer (2005). CODE. Collaborative Ownership in the Digital Economy. Cambridge: MIT Press

[15] Goosen, Frank; Ijntema, Wouter; Frasincar, Flavius; Hogenboom, Frederik; Kaymak, Uzay (2011). News Personalization using the CF-IDF Semantic Recommender. In: Proceedings of WIMS'11, May 25-27, 2011, p. 1-12

[16] Goth, Greg (2006). Data-Driven Enterprise. Slouching toward the Semantic Web. In: IEEE Distributed Systems Online, 7/3, 2006, p. 1-5

[17] Graube, Markus; Pfeffer, Johannes; Ziegler, Jens; Urbas, Leon (2011). Linked Data as integrating technology for industrial data. In: 2011 International Conference on Network-Based Information Systems, 7-9 Sept. 2011, p. 162-167

[18] Haase, Kenneth (2004). Context for Semantic Metadata. In: MM'04, October 10–16, 2004, New York, USA. ACM

[19] Hausenblas, Michael (2009). Exploiting Linked Data to Build Web Applications. In: IEEE INTERNET COMPUTING, 13/4, 2009, p. 68-73

[20] Hee Kyung Lee; Hui Yong Kim; Han-Kyu Lee (2007). News package service based on TV-Anytime metadata gathered from RSS. In: IEEE International Symposium on Consumer Electronics, 20-23 June 2007, p.1-6

[21] Heino, Norman; Tramp, Sebastian; Auer, Sören (2011). ManagingWeb Content using Linked Data Principles – Combining semantic structure with dynamic content syndication. In: 35th IEEE Annual Computer Software and Applications Conference, 18-22 July 2011, p. 245-250

[22] Hoxha, Julia; Brahaj, Armand; Vrandecic, Denny (2011). open.data.al - Increasing the Utilization of Government Data in Albania. In: Proceedings of the 7th International Conference on Semantic Systems, ACM, p. 237-240

[23] Hu, Biyun; Jun Wang; Yiming Zhou (2009). Ontology Design for Online News Analysis. In: WRI Global Congress on Intelligent Systems, 19-21 May 2009, p.202-206

[24] Ijntema, Wouter; Goossen, Frank; Frasincar, Flavius; Hogenboom, Frederik (2010). Ontology-based news recommendation. In: EDBT '10 Proceedings of the 2010 EDBT/ICDT Workshops, 22–26 March 2010, p. 1-6

[25] Jokela, Sami; Turpeinen, Marko; Kurki, Teppo; Savia, Eerika; Sulonen, Reijo. (2001). The role of structured content in a personalized news service. In: System Sciences, 2001. Proceedings of the 34th Annual Hawaii International Conference. p. 1-10

[26] Kim, Hak Lae; Passant, Alexandre; Breslin, John G.; Scerri, Simon; Decker, Stefan (2008). Review and Alignment of Tag Ontologies for Semantically-Linked Data in Collaborative Tagging Spaces. In: IEEE International Conference on Semantic Computing, 4-7 Aug. 2008, p. 315-322

[27] Knauf, Robert; Kürsten, Jens; Kurze, Albrecht; Ritter, Marc; Berger, Arne; Heinich, Stephan; Eibl, Maximilian (2011). Produce. annotate. archive. repurpose --: accelerating the composition and metadata accumulation of tv content. In: Proceedings of the 2011 ACM international workshop on automated media analysis and production for novel TV services, 1 December 2011, p. 31-36

[28] Kobilarov, Georgi; Scott, Tom; Raimond, Yves; Oliver, Silver; Sizemore, Chris; Smethurst, Michael; Bizer, Chris; Lee, Robert (2009). Media meets Semantic Web – How the BBC uses DBpedia and Linked Data to make Connections. In: Proceedings of ESWC 2009, the 6th European Semantic Web Conference. New York: Springer LNCS, p. 723 – 747

[29] Kodama, Masayuki; Ozono, Tadachika; Shintani, Toramatsu; Aosaki, Yasuyoshi (2008). Realizing a News Value Markup Language for News Management Systems Using NewsML. In: International Conference on Complex, Intelligent and Software Intensive Systems, 4-7 March 2008, p.249-255

[30] Kosch, Harald; Boszormenyi, László; Doller, Mario; Libsie, Mulugeta; Schojer, Peter; Kofler, Andrea (2005). The life cycle of multimedia metadata. In: Multimedia, IEEE , vol.12, no.1, p. 80-86

[31] Kurz, Thomas; Schaffert, Sebastian; Bürger, Tobias (2011). LMF - A Framework for Linked Media. In: Workshop on Multimedia on the Web (MMWeb), Sept. 8, 2011, p. 16-20

[32] Larson, Rob; Sandhaus, Evan (2009). NYT to Release Thesaurus and Enter Linked Data Cloud. In: http://open.blogs.nytimes.com/2009/06/26/nyt-to-release-

thesaurus-and-enter-linked-data-cloud/, visited April 20, 2012

[33] Latif, Atif; Us Saeed, Anwar; Höfler, Patrick; Stocker, Alexander; Wagner, Claudia (2009). The Linked Data Value Chain: A Lightweight Model for Business Engineers. In: Proceedings of I-Semantics 2009, the 5th International Conference on Semantic Systems. Graz: Journal of Universal Computer Science, p. 568–577

[34] Liu, Yan; Wang, QingXian; Guo, Lei; Yao, Qing; Lv, Nan; Wang, Qiang (2007). The Optimization in News Search Engine Using Formal Concept Analysis. In: 4th International Conference on Fuzzy Systems and Knowledge Discovery, 24-27 Aug. 2007, p. 45-49

[35] Mannens, E.; Troncy, R.; Braeckman, K.; Van Deursen, D.; Van Lancker, W.; De Sutter, R.; Van de Walle, R. (2009). Automatic metadata enrichment in news production. In: 10th Workshop on Image Analysis for Multimedia Interactive Services, 6-8 May 2009, p.61-64

[36] Messina, Alberto; Montagnuolo, Maurizio; Di Massa, Riccardo; Elia, Andrea (2011). The Hyper Media News System for Multimodal and Personalised Fruition of Informative Content. In: Proceedings of ICMR ’11, April 17-20, 2011, p. 1-2

[37] Ohtsuki, Katsutoshi; Bessho, Katsuji; Matsuo, Yoshihiro; Matsunaga, Shoichi; Hayashi, Yoshihiko (2006). Automatic multimedia indexing: combining audio, speech, and visual information to index broadcast news. In: Signal Processing Magazine, IEEE , vol.23, no.2, p.69-78

[38] Paulheim, Heiko (2011). Improving the usability of integrated applications by using interactive visualizations of linked data. In: Proceedings of the International Conference on Web Intelligence, Mining and Semantics WIMS '11, ACM, p.1-12

[39] Porter, Michael (1985). Competitive Advantage. New York: Free Press

[40] Rayfield, Jem (2012). Sports Refresh: Dynamic Semantic Publishing. In: http://www.bbc.co.uk/blogs/bbcinternet/2012/04/sports_dynamic_semantic.html, visited April 20, 2012

[41] Saumure, Kristie; Shiri, Ali (2008). Knowledge organization trends in library and information studies: a preliminary comparison of pre- and post-web eras. In: Journal of Information Science, 34/5, 2008, p. 651–666

[42] Schandl, Bernhard; Haslhofer, Bernhard; Bürger, Tobias; Langegger, Andreas; Halb, Wolfgang (2011). Linked Data and multimedia: the state of affairs. In: Multimedia Tools and Applications, Online First, p. 1-34

[43] Schouten, Kim; Ruijgrok, Philip; Borsje, Jethro; Frasincar, Flavius; Levering, Leonard; Hogenboom, Frederik (2010). A semantic web-based approach for personalizing news. In: SAC '10 Proceedings of the 2010 ACM Symposium on Applied Computing, 22-26 March 2010, p. 854-861

[44] Smethurst, Michael (2009). BBC Backstage SPARQL endpoint for programmes and music. In: http://www.bbc.co.uk/blogs/radiolabs/2009/06/bbc_backstage_sparql_endpoint.shtml, visited April 20, 2012

[45] Smith, John R.; Schirling, Peter (2006). Metadata standards roundup. In: IEEE Multimedia, 13/2, 2006, p. 84-88

[46] Solomou, Georgia D.; Kalou, Aikaterini K.; Koutsomitropoulos, Dimitrios A.; Papatheodorou, Theodore S. (2011). A Mashup Personalization Service based on Semantic Web Rules and Linked Data. In: 7th International Conference on Signal Image Technology & Internet-Based Systems, Nov. 28 2011-Dec. 1 2011, p. 89-96

[47] Sonntag, Michael (2006). Rechtsschutz für Ontologien. In: Schweighofer, Erich; Liebwald, Doris; Drachsler, Matthias; Geist, Anton (Eds.). e-Staat und e-Wirtschaft aus rechtlicher Sicht. Stuttgart: Richard Boorberg Verlag, p. 418-425

[48] Tesic, Jelena (2005). Metadata practices for consumer photos. In: Multimedia, IEEE , vol.12, no.3, p. 86-92

[49] Yu, Hong Qing; Benn, Neil; Dietze, Stefan; Pedrinaci, Carlos; Liu, Dong; Domingue, John; Siebes , Ronald (2010). Two-staged approach for semantically annotating and brokering TV-related services. In: IEEE International Conference on Web Services, 5-10 July 2010, p. 497-503

[50] Zimmermann, Antoine (2011). Leveraging the Linked Data Principles for Electronic Communications. In: IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, 22-27 Aug. 2011, p. 385-388

[51] Zhou, Yan-quan; Hu, Ying-fei; He, Hua-can (2007). Learning User Profile in the Personalization News Service. In: International Conference on Natural Language Processing and Knowledge Engineering, 2007, p.485-490

[52] Pellegrini, Tassilo (2009). Kollaborative Strategien der Klassifizierung von AV-Content. In: Krone, Jan (Hg.). Fernsehen im Wandel. Mobile TV & IT-TV in Deutschland und Österreich. Baden-Baden: Nomos, S. 107 – 132

[53] Fine, Nick; Brinkman, Willem-Paul (2004). Informing Intelligent Environments: Creating Profiled User Interfaces. In: Proceeings of EUSAI 2004, November 8-10, p. 14-18

[54] Malheiros, Miguel; Jennett, Charlene; Patel, Snehalee; Brostoff, Sacha; Sasse, M. Angela (2012). Too Close for Comfort: A Study of the Effectiveness and Adaptability of Rich-Media Personalized Advertising. In: Proceedings of CHI 2012, May 5–10, 2012, p. 579-588

[55] Lane, Nicholas D.; Hong Lu, Ye Xu; Campbell, Andrew T.; Choudhury, Tanzeem; Eisenman, Shane B. (2011). Exploiting Social Networks for Large-Scale Human Behavior Modeling. In: IEEE Pervasive Computing, October-November 2011, p. 45-53

[56] Rayfield, Jem (2010). BBC World Cup 2010 dynamic semantic publishing. In: BBC Internet Blog, http://www.bbc.co.uk/blogs/bbcinternet/2010/07/bbc_world_cup_2010_dynamic_sem.html, visited May 5, 2012

[57] Rayfield, Jem (2012). Sports Refresh: Dynamic Semantic Publishing. In: BBC Internet Blog, http://www.bbc.co.uk/blogs/bbcinternet/2012/04/sports_dynamic_semantic.html, visited May 5, 2012