Linked Data Utilization along the Content Value Chain – Observations and Implications

Georg Neubauer
University of Applied Sciences St. Poelten
Matthias Corvinus Str. 15, 3100 St. Poelten, Austria
dm131520@fhstp.ac.at

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

ABSTRACT
The authors present the results of a longitudinal investigation in the utilization of Linked Data technologies along the content value chain. The authors analyzed 71 papers in the period from 2006 to 2014 that used Linked data technologies in editorial workflows. By coding the primary and secondary research topics addressed in the paper the authors draw a conclusion of the maturity of Linked Data technologies as support systems along the content value chain. The survey indicates that Linked Data technologies are constantly maturing as a support infrastructure for editorial processes. The validity of the survey results for application domains not related to editorial tasks is open to discussion.

Categories and Subject Descriptors
E.0 [**General**]; K.4.3 [**Organizational Impacts**]

General Terms
Management, Economics, Human Factors, Standardization

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

1. INTRODUCTION
The growing recognition of Linked Data among the research community as "Semantic Web done right" [14] motivates to take a closer look if and how Linked Data research has evolved over the recent years. Such an investigation allows to gain insights into research trends and interdependencies thereof, and it allows to draw conclusions whether the research field has reached a significant degree of maturity in terms of technology diffusion and application areas.

As illustrated in Figure 1 a survey about the occurrence of the phrase "Linked Data" in research publications of the ACM digital library from the period 2006 to 2014 reveals the growing popularity of this technological concept in the computer sciences till 2013 with a decline in 2014. Linked Data as a generic technology for data management is being applied across various application areas and industries, making it very hard to come to a general statement concerning its level of maturity and industry adoption. So is this distribution from figure 1 an indicator for the growing maturity of a research field? And if yes, how can this maturity be operationalized empirically?

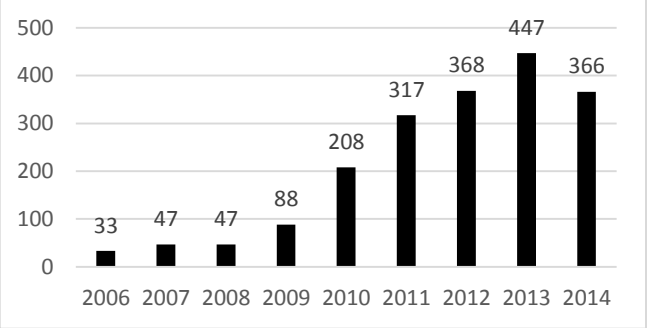

Figure 1. ACM Publications containing the term "Linked Data" from 2006 – 2014 (N = 1921)

To tackle these questions the authors chose to analyze a subset of research papers from the ACM database that address the application of Linked Data within editorial workflows. This subset allowed us to apply a unified classification scheme – known as the content value chain [1] – to the various application areas of Linked Data. The content value chain can be described as a process model that is comprised of several sequential steps contributing to the content production process. By looking at the application area of Linked Data in editorial workflows it was possible to identify primary and secondary areas of utilization, thus allowing us to draw conclusions towards the diffusion and appropriability of Linked Data for the production of media content.

2. CLASSIFICATION SCHEME & RELATED WORK
The original concept of the value chain as developed by Michael Porter in 1979 is used as an analytical framework for the analysis of value creation processes at the firm level or the industry level [15]. Over recent years the concept of the value chain has also gained popularity in the context of open data in general [4; 6; 16] and Linked Data in special [3; 5]. Especially research that investigated the organizational and economic impact of Linked Data refers to the concept of the value chain [13].

In this paper we refer to a generic abstraction of the content value chain consisting of five steps: 1) content acquisition, 2) content editing, 3) content bundling, 4) content distribution and 5) content consumption. As illustrated by [1] Linked Data can contribute to each step by supporting its associated intrinsic production function. These are in detail:

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.

Content editing 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.

Content bundling is mainly concerned with the contextualization and personalization 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 content, or to compile thematically relevant material into Landing Pages or Dossiers thus improving the navigability, findability and reuse of information.

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. These can be designed either to serve internal purposes so that data can be reused within controlled environments (i.e. within or between units) or for external purposes so that data can be shared between unknown users (i.e. as open SPARQL Endpoints on the Web).

Content consumption 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 content i.e. by providing reasonable retrieval tools and/or visualizations.

The five steps of the content value chain comprise the classification scheme.

3. METHODOLOGY

We selected a sample of 71 papers (out of 1921) dealing with the utilization of Linked Data in editorial workflows in the period from 2006 to 2014 from the ACM Digital Library (DL). The selected papers had to comply with the following criteria: 1) the work must analyse the utilization of Linked Data with reference to some sort of editorial workflow; and 2) the work must not be purely theoretical but provide at least a proof of concept. The relevant papers have then been analysed and clustered according to the five classes acquisition, editing, bundling, distribution, consumption. As most papers treated more than one of these topics we weighted each paper according to the primary and secondary topic discussed, thus also gaining a better understanding how the research topics relate to each other.

Figure 2 illustrates the classification scheme. The black boxes indicate the primary classification of a paper and the amount of papers falling into this category. The secondary classification inherit a weighted greyscale value. The number in the grey and black boxes indicates how many papers referred to these classes. Hence, reading the rows horizontally gives an overview how the primary classification of a paper relates to its secondary classification. Reading the columns vertically by summing up the values from the black boxes gives the amount of papers falling into a specific class.

The weighted greyscale values have been calculated as follows. Given that black is 100%. 50% divided by the amount of papers with main classification (black) multiplied with the amount of the related classifications for the secondary classification. Figure 4 illustrates the results of our survey.

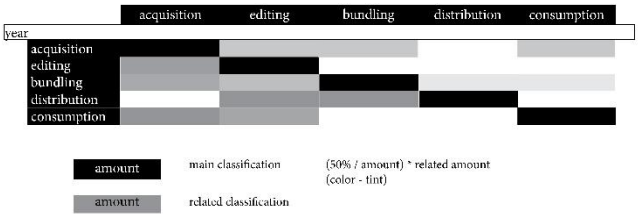

Figure 2. Legend: time-based categorization into the content value chain

4. RESULTS

4.1 General Findings

Figure 3 illustrates the general findings of our investigation, which are showing the result of all years later discussed in 4.2 as influence circles on a grid. The diagonal line with the black circles represent the amount of papers within the main classification, while the other circles show the related classifications if they are read in a horizontal way. As mentioned, related classifications are a result of additionally found secondary topics that match the content value chain, for one paper already has a main topic classification.

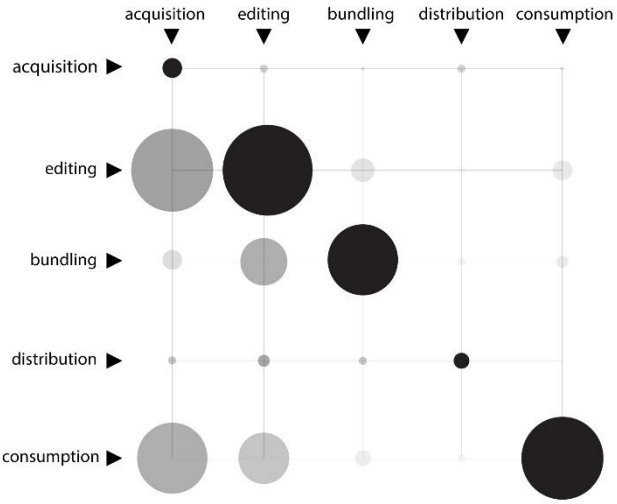

Figure 3. Influence cycles (result) – time-based categorization into the content value chain

The main application areas of Linked Data in editorial workflows fall into the areas editing (23 papers), bundling (18 papers) and consumption (21 papers).

Crawling and leveraging processes could be subsumed as acquisition process [1] using special indexing methods for several entities found and aggregated through queries. The indexing methods built a fundament for further scientific processing called content editing.

Scientific editing using algorithmic methods to classify data into separated, semantically enriched lists or ontologies were treated in

23 papers as main topic. All of these editing methods were part of a recognition process used for video-, text- or graphic- analysis in terms of media-analysis and enrichment of metadata.

18 papers concerned content bundling as main topic. Bundling can easily be defined as fine-grained representations of resource parts used for personalization and contextualization of the content.

Just 4 papers described distributions for example in case of improved accessibility of information. The main difference to the content bundling process and the content consumption process explained later on, therefore was, that only APIs can access this data which in case of content bundling wasn't put to visualized graphs of the content. This low number of distributions is not significant for further conclusions.

21 papers applied Linked Data through a framework visualizing graph-based relations of links. This sort of standard for framework developers was to visualize links of Linked Data for purposes like content recommendation.

4.2 Longitudinal Perspective

Figure 4 illustrates the results of our analysis from a longitudinal perspective. The visualization scheme corresponds with Figure 3 but additionally lists the amount of papers (the black boxes) and their related topics (the grey boxes) in the years from 2006 to 2014. I.e. if there are two papers of content acquisition in 2014, this means that these two papers have their main classification in content acquisition and related topics in all other areas of the value chain.

2006: We found just one paper in 2006 with relation to our research focus. This paper addressed content acquisition as main topic and editing issues as secondary topic.

2007: In 2007 one paper was classified treating content bundling as main topic and content acquisition as secondary topic. Two papers addressing content consumption as primary topic and acquisition, editing and bundling in treating only content consumption.

2008: In 2008 we determine one paper addressing content distribution and one paper addressing content consumption both referring to content editing.

2009: We have three papers classified as content editing, content bundling and content consumption. The subrelations in case of content bundling is editing and in case of content consumption the subrelations equally refer to content bundling and content distribution.

2010: In 2010 the authors detected one paper treating content acquisition, one paper treating content distribution and another one content consumption. Two papers treated content editing frameworks. All of the five papers treated content acquisition as their secondary topic.

2011: In 2011 one paper was about content acquisition, editing, distribution and content consumption. The relations begin in the content editing class including a single subrelation to content acquisition and content consumption. Four papers have all an equal amount of subrelations to content acquisition and editing. Additionally one paper described a framework for content consumption.

2012: In 2012 the authors found one paper addressing content acquisition as main topic and content editing as secondary topic. Two papers demonstrated the opposite pattern, discussing editing as main topic and acquisition as secondary topic. Four papers refer to content bundling with subrelations to content acquisition and content editing, while one of them also mentioned content distribution or content consumption as tertiary topic. Four papers address content consumption as main topic showing subrelations to content acquisition in all of their descriptions and one paper including further treatment of editing.

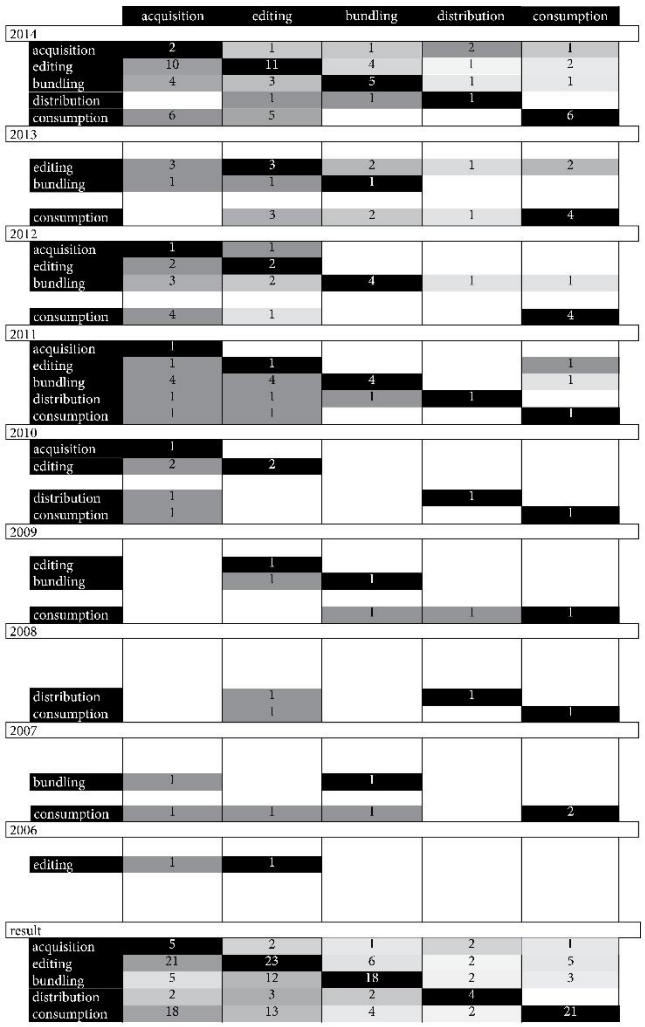

Figure 4. Primary and secondary topics in Linked Data utilization

2013: All papers that describe content editing frameworks in the year of 2013 also have acquisitional processes as topic. One of three papers addressing content editing have a subrelation to content bundling. Two papers are subrelated to content distribution and one to content consumption. Only one paper related to content bundling subrelated to content acquisition and content editing. Four papers give reason to content consumption. Their relation to subclasses are three addressing content editing, two addressing content bundling and four addressing content consumption frameworks as main topic.

2014: In 2014 the classification scheme of the content value chain seems applicable to a huge amount of papers. We analysed 25 papers and came to the conclusion that scientific content editing utilizing combinations of vocabularies for the preparation of linked data is high of note, i.e. automatic extraction RDF-Triples from web sources for purposes of content enrichment. So 11 papers are classified as content editing in nearly all cases within

acquisitional preprocessing. Content bundling with 5 papers and content consumption with 6 papers as main classification seem very similar spreaded in relation to the former years.

5. DISCUSSION, LIMITATIONS & FUTURE WORK

The results show a trend in the utilization of Linked Data technologies towards content editing, content bundling and content consumption. Especially the increasing amount of papers addressing consumption purposes after 2009 is taken as an indicator for the increasing maturity of Linked Data technologies in editorial workflows. We also made out a reason of the increasing usage of content acquisition processes beginning in 2008, assuming that the data infrastructure achieved reclaimable integrity. Concerning the main result the intertwinedness of research topics have seamless integration of distinct steps in the content value chain. Metadata acquisition systems can minimize the human burden in recording data [12]. Normally the content acquisition process is the premier step to process data. We also claim that there exists a structural relation between content distribution and acquisition given the fact that these two processes are technologically intertwined in interlinked data ecosystems. Content distribution could be treated as a main goal of data storage and supply [13]. The authors assume that well established Linked Data stores are a precondition to content acquisition allowing further processing like content bundling, content distribution and content consumption. By taking this appropriate amount of papers in 2014 we came to the conclusion that content editing takes root, but the consistency of the result should also be considered in a normalized way to the former years.

To gain further insights the authors plan to extend the sample size of their survey in their future work. The current amount of 71 papers is simply too small to draw precise conclusions on the state of the art and future direction of Linked Data utilization in editorial workflows. But apart from these limitations the insights generated by the survey indicate that Linked Data technologies are constantly maturing as a support infrastructure for editorial processes. The validity of the survey results for application domains not related to editorial tasks is open to discussion.

6. REFERENCES

[1] Pellegrini, Tassilo. “Integrating Linked Data into the Content Value Chain: A Review of News-Related Standards, Methodologies and Licensing Requirements.” In Proceedings of the 8th International Conference on Semantic Systems, 94–102. ACM, 2012. http://dl.acm.org/citation.cfm?id=2362513.

[2] Auer, Sören, Theodore Dalamagas, Helen Parkinson, François Bancilhon, Giorgos Flouris, Dimitris Sacharidis, Peter Buneman, et al. “Diachronic Linked Data: Towards Long-Term Preservation of Structured Interrelated Information.” In *Proceedings of the First International Workshop on Open Data*, 31–39. WOD ’12. New York, NY, USA: ACM, 2012. http://doi.acm.org/10.1145/2422604.2422610.

[3] Auer, Sören, Jens Lehmann, Axel-Cyrille Ngonga Ngomo, and Amrapali Zaveri. “Introduction to Linked Data and Its Lifecycle on the Web.” In *Reasoning Web. Semantic Technologies for Intelligent Data Access*, 1–90. Springer, 2013. http://link.springer.com/chapter/10.1007/978-3-642-39784-4_1.

[4] Davis, Mills. “The Business Value of Semantic Technologies.” Presentation and Report, Semantic Technologies for E-Government, 2004. http://project10x.com/bio_downloads/business_value_of_semanti c_technologies_2005.pdf, accessed May 9, 2015

[5] Latif, Atif, Anwar Us Saeed, Patrick Hoefler, Alexander Stocker, and Claudia Wagner. “The Linked Data Value Chain: A Lightweight Model for Business Engineers.” In I-SEMANTICS, 568–75. Citeseer, 2009. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.181.950 &rep=rep1&type=pdf.

[6] Pepe, Alberto, Matthew Mayernik, Christine L. Borgman, and Herbert Van de Sompel. “Technology to Represent Scientific Practice: Data, Life Cycles, and Value Chains.” World Wide Web Internet And Web Information Systems, 2009, 1–22.

[7] Robak, Silva, Bogdan Franczyk, and Marcin Robak. “Research Problems Associated with Big Data Utilization in Logistics and Supply Chains Design and Management,” n.d. https://fedcsis.org/proceedings/2014/pliks/472.pdf.

[8] Solanki, Monika, and Christopher Brewster. “Consuming Linked Data in Supply Chains: Enabling Data Visibility via Linked Pedigrees.” In COLD, 2013. http://windermere.aston.ac.uk/~monika/papers/SolankiCOLD2013.pdf.

[9] Taskar, Benjamin, Eran Segal, and Daphne Koller. “Probabilistic Classification and Clustering in Relational Data.” In International Joint Conference on Artificial Intelligence, 17:870–78. LAWRENCE ERLBAUM ASSOCIATES LTD, 2001. http://ai.stanford.edu/users/koller/Papers/Taskar+al:IJCAI01.pdf.

[10] Van Erp, Marieke, Willem Robert van Hage, Laura Hollink, Anthony Jameson, and Raphaël Troncy. “Detection, Representation, and Exploitation of Events in the Semantic Web,” 2013. http://ceur-ws.org/Vol-1123/proceedingsderive2013.pdf.

[11] Villazón-Terrazas, Boris, and Oscar Corcho. “Methodological Guidelines for Publishing Linked Data.” *Una Profesión, Un Futuro: Actas de Las XII Jornadas Españolas de Documentación: Málaga* 25, no. 26 (2011): 20.

[12] Labrinidis, Alexandros, and H. V. Jagadish. “Challenges and Opportunities with Big Data.” *Proc. VLDB Endow.* 5, no. 12 (August 2012): 2032–33. doi:10.14778/2367502.2367572.

[13] Edward, Curry et al. "Big Data. Technical Working Groups White Paper," 2014. http://bigproject.eu/sites/default/files/BIG_D2_2_2.pdf

[14] Berners-Lee, Tim (2008). Linked open Data. See also: http://www.w3.org/2008/Talks/0617-lod-tbl/#%281%29, accessed May 9, 2015

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

[16] Archer, Phil; Dekkers, Max; Goedertier, Stijn; Loutas, Nikolaos (2013). Study on business models for Linked Open Government Data (BM4LOGD - SC6DI06692). Services See also: http://ec.europa.eu/isa/documents/study-on-business-modelsopen-government_en.pdf, accessed May 10, 2015