Towards Semantic Story Telling with Digital Curation Technologies

Julian Moreno Schneider, Peter Bourgonje, Jan Nehring, Georg Rehm, Felix Sasaki, Ankit Srivastava
DFKI GmbH, Language Technology Lab, Alt-Moabit 91c, 10559 Berlin, Germany
dkt@dfki.de

Abstract

We develop a system that aims at generating stories or, rather, potential story paths, based on the semantic analysis of multiple source documents (including news articles) using template-filling. The final system will be realised by additional methods, also taking specific domains and topics into account. For the processing we use NLP methods such as named entity recognition, we also use a triple store and classic document indexing modules. The analysis information is filtered, rearranged and recombined to fit the respective template. The system's use case is to support knowledge workers (journalists, editors, curators etc.) in their tasks of processing large amounts of (incoming) documents, to identify important entities, relationships between entities and to suggest individual story paths between entities, eventually to come up with more efficient processes for content curation.

1 Introduction and Context

Journalists have to cope with huge amounts of incoming information that need to be scanned, skimmed, contextualised, evaluated and, eventually, further processed into a new piece of news, blog post, or longer article. The demand for tool support is extremely high. Many journalists and online creators are under a lot of pressure as they are expected to produce as many pieces as possible in less and less time.

However, it is not only journalists who have a growing demand for semantic tools to help them with data processing (in terms of efficiency, breadth, depth, scope etc.), ascertaining what is important and relevant, maybe even genuinely new, surprising and eye-opening. In addition to journalists who work for traditional or online news outlets (incl. blogs, newspapers, radio and tv stations etc.), there are many other job profiles that have to cope with a rather high volume of incoming news or, on a more general level, content that need to be processed in a rather short amount of time in order to produce something new – let us call this group “knowledge workers”. These can be, among others, authors and creatives who work in an agency specialised on building information portals: the client provides a smaller or larger amount of data, information, documents, and pictures that now needs to be processed into an interactive website. Second example: creatives who work in an agency that specialises on conceptualising and producing museum exhibitions and showrooms. On a regular basis, these teams face the challenge of becoming experts on a completely new topic basically overnight, when they are confronted with a huge pile of highly domain-specific information that needs to be transformed into an exhibition (or into a convincing pitch to actually get the contract for the production of the new exhibition).

The common ground of the different tasks and challenges described above is the *curation of digital content*. In our project Digital Curation Technologies[1] we collaborate with four SME companies that cover the different use cases and sectors mentioned above, including journalism [Rehm and Sasaki, 2016]. The goal of our project is to design and to build language and knowledge technologies that support the knowledge workers and that help them to become more efficient by delegating routine tasks to the machine with a focus on use-case specific text documents (we currently work on data sets provided by our four SME partners) so that the knowledge workers can concentrate on their core tasks, i. e., producing a story or document that is based on a specific genre or text type (a news piece, an exhibit, a tv news report etc.) and that relies on facts and figures contained in a heterogeneous collection of content.

Among the tools that we develop and integrate into our emerging Platform for Digital Curation Technologies are semantic story telling, named-entity recognition, entity linking, temporal analysis, machine translation, summarisation, classification and clustering [Bourgonje *et al.*, 2016]. We currently focus upon providing RESTful APIs to our SME partners that provide basic functionalities that can already now be integrated into their own in-house systems. In addition, we work on the more complex, longer-term idea of designing and implementing a system for Semantic Story Telling. This system will eventually be able to take a large amount of documents, extract entities and relations between entities, also extract temporal information and relationships, automatically produce a hypertext view of the document cluster in order to enable knowledge workers quickly and efficiently to familiarise themselves with the document collection (i. e., with a new domain or a new topic). We also experiment with the

[1] DKT, see http://www.digitale-kuratierung.de for more details.

idea of automatically generating story paths through this hypertext cluster that can then be used as the foundation of a new piece of content. In a later stage of the project we plan to augment our technologies with state of the art big data systems in order to be able to process high volumes of news data in motion.

The paper is structured as follows: Section 2 discusses related work. Section 3 presents the current architecture of our system. An initial evaluation is described in Section 4. Section 5 concludes the article.

2 Related Work

Our Semantic Story Telling approach is rooted in and influenced by several different approaches in the area of text understanding and generation. [Rumelhart, 1975] was among the first to break down texts of specific types into smaller components by introducing the notion of story grammars that provide established conventions with regard to structure, contents and expectations. Multiple authors developed these concepts further, see, for example, [Orlikowski and Yates, 1994], who define a genre as "a distinctive type of communicative action, characterised by a socially recognised communicative purpose and common aspects of form. The communicative purpose of a genre is not rooted in a single individuals motive for communicating, but in a purpose that is constructed, recognised, and reinforced within a community". [Mann and Thompson, 1988] introduced Rhetorical Structure Theory as a means of describing and specifying the rhetorical relationships between parts of a text. [Rehm, 2002] combined several of these approaches into a system for the automatic identification of different web genres. [Rehm, 2005] provides a comprehensive overview of the literature.

In our current, early prototype implementation we combine several different NLP and IR methods with the goal of providing curation technologies to knowledge workers in different sectors. While many approaches, for example, in the context of museums or libraries, focus upon digital museums (i. e., form and presentation) [Y.-C. Li, 2012] or digital libraries [Meghini and Bartalesi, 2014], we focus upon supporting the actual internal procedures and processes that are used for preparing a real or digital museum (through semantic analysis of the content to be curated, automatically creating metadata etc.). We concentrate on the discoverability of curated content and establishing semantic relations between concepts (i. e., entities or relations) to improve understanding of the subject of research. We also include external ontologies and linked data sources. The system described by [Lewis *et al.*, 2014] is similar to our approach but it is targeted at the localisation industry and uses different data models.

3 System Overview

The current Digital Semantic Storytelling System (DS3) prototype involves several components and two main processing phases (see Figure 1). The user manually selects a story template (Section 3.1) that is then processed and filled (Section 3.2). This process is based on a semantic layer [Bourgonje *et al.*, 2016] that is constructed on top of the respective document collection (Section 3.3).

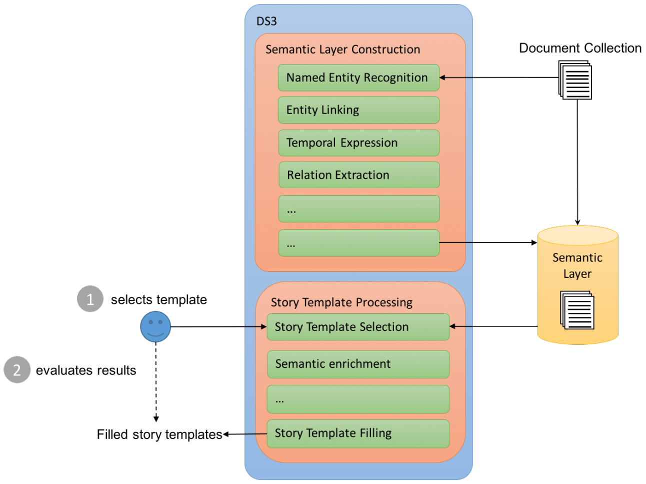


Figure 1: The architecture of the DS3 prototype

3.1 Template Definition and Selection

The process starts with the user selecting the template that applies to and best fits the particular topic and domain of the document collection. The prototype currently contains two templates. Templates are defined in a Template Pattern File (TPT). A TPT file contains the title of the template followed by information about its structure and content (the individual fields of the template). A field definition consists of several columns, the number of columns depends on the type of the field (defined in the second column). A missing value in a column is represented by a single question mark. We currently define the following columns:

1. **Name:** the name of this field
2. **Type:** the type of the field; currently there are two possible field types: single (an entity with an associated URL) and triple (used for storing relations)
3. **Required:** boolean (field required: yes/no)
4. **Subject:** required if field type is triple
5. **Predicate:** required if field type is triple
6. **Object:** required if field type is triple

For the full system we need to significantly extend not only the conceptual specification of templates but also the currently implemented set of templates. For now we have included two templates (*Biography* and *News*). In the following, we focus upon the News template.

- **Biography:** *maincharacter*, *dateofbirth*, *placeofbirth*, *dateofdeath*, *placeofdeath*, *placeofresidence*, *relationship* (*maincharacter* field is single type, the other fields are of type triple).
- **News:** *mainfact*, *locations*, *persons*, *organisations*, *times/dates* (*mainfact* field is single type, the other fields are of type triple).

3.2 Template Filling

The template defines – in the current prototype still in a rather loose sense – the structure of a story. In the Template Filling phase, analysis information provided by the Semantic Layer (see Figure 1) is used to fill the template. In DS3 we use the Sesame framework for semantic storage. The Semantic Layer contains information that originates in external ontologies (DBPedia, German National Library, verb ontologies etc.) as well as several NLP methods.

3.3 Semantic Layer Construction

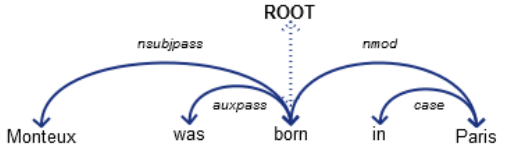
The semantic analysis consists of a pipeline that combines named entity recognition (NER), entity linking, temporal analysis and simple relation extraction. The modules in the pipeline are connected through the NLP Interchange Format (NIF) [Sasaki *et al.*, 2015]. Each analysis takes either plain text or NIF as input and outputs NIF, in which the additional semantic information is stored as annotations.

Our NER module is based on OpenNLP. The approach combines models (if training data is available) and dictionaries (if domain-specific data such as compiled word lists or gazetteers, provided by our SME partners, is available). For every recognised entity we attempt to retrieve a URI in either a domain-specific ontology or DBPedia. If we successfully retrieve a URI, we proceed to use type-specific SPARQL queries to retrieve additional information (e. g., latitude and longitude points for locations, date of birth and death for persons).

The Temporal Analysis module is based on a regular expression grammar. Recognised expressions are resolved to a fully-specified format. For underspecified dates, an anchor date is used for normalisation. This is either the creation date of the document (if available) or another, previously normalised temporal expression. The final annotation is always a range, allowing the inclusion of more specific dates in less specific dates (i. e., we can recognise that, e. g., "13-04-2015" is part of "April of 2015"). The module is rule-based and currently works for English and German.

Explicitly encoded or annotated relations are an important prerequisite for attempting to generate story paths over a set of concepts or entities. In terms of relation extraction we currently experiment with the Stanford CoreNLP dependency parser [Manning *et al.*, 2014]. Our current approach is to extract subject-relation-object triples for which the governing node is a verb. The dependency that has a subject type relation is taken as the subject and the dependency having an object type relation is taken as the object. We subsequently filter for triples for which the subject and the object are named entities for which a URI has been retrieved. These are stored in an internal ontology. This results in a collection of relation triples that we can use to fill templates. Figure 2 shows an example dependency graph for the sentence "Monteux was born in Paris" and the corresponding NIF annotation; by using token indices we can combine the two to arrive at the triple *[http://d-nb.info/gnd/122700198, born, http://www.geonames.org/2988507]*, which can fill the birth place slot in a biography template. A drawback of this approach we found is that the number of relations that were extracted are relatively limited. This is due to the requirement that both the subject and the object of the triple must be governed by the same verb node, and the filtering step, where we keep only those relation triples for which both the subject and the object was and could be recognized as an entity and also resolved to a URI. We want to ensure connectability with external ontologies, thus keep the filtering in place (e. g., a relation triple like *[Mary, met, John]* where we have no further information regarding *Mary* or *John* in the form of a URI in an ontology is currently only of limited use for our application). To increase the number of useful relations in future versions of our relation extraction component, we are experimenting with finding the lowest governing node that is a verb that connects the subject and object nodes in the graph. This verb will then be taken as the type of relation. In addition, we will look into other, dedicated and more sophisticated tools and approaches for relation extraction. Furthermore, not only do we have to identify relations between entities or concepts and their specific features or individual characteristics, but also relations with regard to, among others, coreference of entities, relations between different parts of a document, relations between the same instances of concepts mentioned in multiple documents, to name just a few.

The semantic annotations mentioned above constitute the semantic layer on top of the document collection that is being processed when filling a template.



```
<http://dkt.dfki.de/documents/#char=0,25>
 a                        nif:RFC5147String , nif:String , nif:Context ;
 nif:beginIndex           "0"^^xsd:nonNegativeInteger ;
 nif:endIndex             "25"^^xsd:nonNegativeInteger ;
 nif:isString             "Monteux was born in Paris"^^xsd:string .

<http://dkt.dfki.de/documents/#char=20,25>
 a                        nif:RFC5147String , nif:String ;
 nif:anchorOf             "Paris"^^xsd:string ;
 nif:beginIndex           "20"^^xsd:nonNegativeInteger ;
 nif:endIndex             "25"^^xsd:nonNegativeInteger ;
 nif:entity               <http://dkt.dfki.de/ontologies/nif#LOC> ;
 nif:referenceContext     <http://dkt.dfki.de/documents/#char=0,25> ;
 itsrdf:taIdentRef        <http://www.geonames.org/2988507> .

<http://dkt.dfki.de/documents/#char=0,7>
 a                        nif:RFC5147String , nif:String ;
 nif:anchorOf             "Monteux"^^xsd:string ;
 nif:beginIndex           "0"^^xsd:nonNegativeInteger ;
 nif:endIndex             "7"^^xsd:nonNegativeInteger ;
 nif:entity               <http://dkt.dfki.de/ontologies/nif#PER> ;
 nif:referenceContext     <http://dkt.dfki.de/documents/#char=0,25> ;
 itsrdf:taIdentRef        <http://d–nb.info/gnd/122700198> .
```

Figure 2: Dependency graph for "Monteux was born in Paris" and the corresponding NIF document

An initial proof-of-concept of the DS3 semantic information retrieval module is available online.[2] Figure 3 shows the filled in *News* story template that was generated for the user-selected initial concept "Erich Mendelsohn". The system provides concepts related to the initial concept. For each related entity, subtrees are built. Our goal with this approach and system is, ultimately, to provide a tool that knowledge workers can use not only to explore a semantic space in an interactive way but to get support for the identification of interesting

[2] http://dev.digitale-kuratierung.de/ds3/

story paths in a potentially huge concept space that the knowledge worker is not familiar with. We will also include a feedback mechanism so that the user can up-vote or down-vote extracted entities and relations. Once the semantic concept space, interactively adjusted and partially ranked by the user, is complete, the selected story path can be exported into the desired format for further processing. This functionality will be implemented in collaboration with our SME partners and tailored to their respective in-house systems.

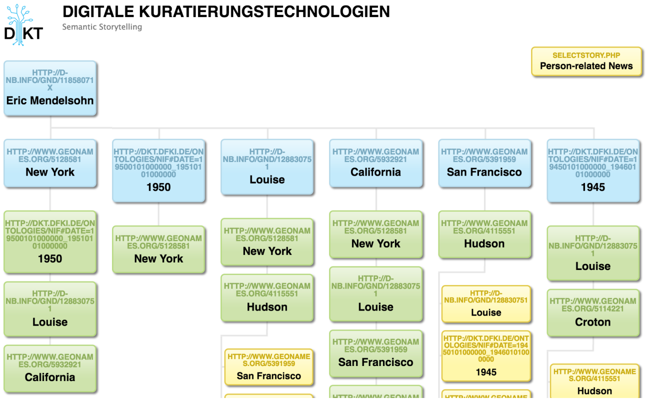

D KT
DIGITALE KURATIERUNGSTECHNOLOGIEN
Semantic Storytelling
SELECTSTORY.PHP
Person-related News
HTTP://D-NB.INFO/GND/11858071X
Eric Mendelsohn
HTTP://WWW.GEONAMES.ORG/5128581
New York
HTTP://DKT.DFKI.DE/ONTOLOGIES/NIF#DATE=19500101000000_19510101000000
1950
HTTP://D-NB.INFO/GND/128830751
Louise
HTTP://WWW.GEONAMES.ORG/5932921
California
HTTP://WWW.GEONAMES.ORG/5391959
San Francisco
HTTP://DKT.DFKI.DE/ONTOLOGIES/NIF#DATE=19450101000000_19460101000000
1945
HTTP://DKT.DFKI.DE/ONTOLOGIES/NIF#DATE=19500101000000_19510101000000
1950
HTTP://WWW.GEONAMES.ORG/5128581
New York
HTTP://WWW.GEONAMES.ORG/5128581
New York
HTTP://WWW.GEONAMES.ORG/5128581
New York
HTTP://WWW.GEONAMES.ORG/4115551
Hudson
HTTP://D-NB.INFO/GND/128830751
Louise
HTTP://D-NB.INFO/GND/128830751
Louise
HTTP://WWW.GEONAMES.ORG/4115551
Hudson
HTTP://D-NB.INFO/GND/128830751
Louise
HTTP://D-NB.INFO/GND/128830751
Louise
HTTP://WWW.GEONAMES.ORG/5114221
Croton
HTTP://WWW.GEONAMES.ORG/5932921
California
HTTP://WWW.GEONAMES.ORG/5391959
San Francisco
HTTP://WWW.GEONAMES.ORG/5391959
San Francisco
HTTP://DKT.DFKI.DE/ONTOLOGIES/NIF#DATE=19450101000000_19460101000000
1945
HTTP://WWW.GEONAMES.ORG/4115551
Hudson

Figure 3: Filled in *News* story template relating to the root concept "Erich Mendelsohn"

4 Use Case and Experiments

The development of the DS3 system is work in progress. In the following we describe an experiment that relates to a typical future use case of the system. Since a key requirement of the whole Digital Curation Technologies project is adaptability to new and specialised domains, we made several experiments with the Mendelsohn Letters, a large set of letters written by Erich Mendelsohn, a well-known German architect.[3] With the current DS3 prototype, we can extract required information and fill templates by using DBPedia in combination with template-specific SPARQL queries. However, for smaller and more specialised domains such a (complete) ontology may not be available. For the Mendelsohn experiments we adapted our NER module to this domain and created semantic annotations for a random sample of 1,000 letters. We extracted the relation triples and evaluated if they are suitable for filling the selected templates. Given our currently still limited set of templates, the amount of information obtained was also limited. We were able to extract several relations from the data set, but only some were useful for filling slots in the templates due to the limited recall of the dependency-based relation extraction (see Section 3.3). In an attempt to extract additional relation candidates, we assumed entities to be related if they appear near each other. Of the window sizes we tried (within 5, 10 and 20 words of each other), we found that a size of 20 gave the best results. This is also what we used to generate the tree shown in Figure 3, where any entities that appear more than ten times together in the same window are shown. This rather coarse-grained approach of finding potentially related entities serves as an alternative, only to demonstrate what our current output looks like. Not in the least place because this proximity approach would only establish the subject and object of the relation triples we use and not the relation type itself, which is taken from the verb through the dependency parsing approach. Finding more informative relation triples using more general and more robust relation extraction approaches with a bigger coverage will be an important next step. Because the individual components we use in the platform (except for the temporal analyser) are typical off-the-shelf implementations with limited modifications, we do not provide F scores for these components. Instead, the focus is on combining existing technologies within a larger platform for Digital Curation Technologies, especially with regard to Semantic Story Telling. As a next step we will do an evaluation in which we will ask a group of knowledge workers to compare their workflow with and without using our tools. A key principle of the project is that the human expert is always in the loop. This means that the performance of individual components is secondary to the efficiency and usability of the services and platform as a whole and evaluation should be user-oriented.

5 Summary and Future Work

We are developing a system that will support knowledge workers in the complex and time-consuming task of handling, evaluating, processing, sorting and processing of document collections – either data in motion coming in, among others, from online news wires, or highly specialised, self-contained document collections. The primary goal of the system is to enable journalists, editors, authors, i. e., curators of digital content to identify interesting story lines as efficiently as possible. The current prototype is able to fill manually selected story templates based on semantically processing a document collection through, for example, named entity recognition and relation extraction. Future work includes the implementation of additional semantic analysis modules (e. g., entity recognition with higher recall through classic IR methods such as TF/IDF, additional as well as template-specific relation extraction methods, exploiting ontologies to make better use of identified relations), more detailed template descriptions, additional templates and crosslingual capabilities through machine translation. Within the context of the project Digital Curation Technologies we plan to test the system in two newsrooms (newspaper, television station) and also in a digital agency that specialises on designing and curating online portals for cultural archives and heritage information.

Acknowledgments: The authors would like to thank the anonymous reviewers for their helpful comments and suggestions. The project Digitale Kuratierungstechnologien is supported by the German Federal Ministry of Education and Research, Unternehmen Region, Wachstumskern-Potenzial (No. 03WKP45). The project FREME has received funding from the EU's Horizon 2020 programme under grant agreement No. 644 771.

[3] http://ema.smb.museum/en/home/

References

[Bourgonje *et al.*, 2016] P. Bourgonje, J. Moreno-Schneider, J. Nehring, G. Rehm, F. Sasaki, and A. Srivastava. Towards a Platform for Curation Technologies: Enriching Text Collections with a Semantic-Web Layer. In H. Sack, G. Rizzo, N. Steinmetz, D. Mladeni, S. Auer, and C. Lange, editors, *The Semantic Web: ESWC 2016 Satellite Events*, June 2016. In print.

[Lewis *et al.*, 2014] D. Lewis, A. Gómez-Pérez, S. Hellmann, and F. Sasaki. The role of linked data for content annotation and translation. In *Proc. of 2014 European Data Forum*, 2014.

[Mann and Thompson, 1988] W. C. Mann and S. A. Thompson. Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. *Text*, 8:243–281, 1988.

[Manning *et al.*, 2014] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky. The Stanford CoreNLP natural language processing toolkit. In *ACL System Demonstrations*, pages 55–60, 2014.

[Meghini and Bartalesi, 2014] C. Meghini and V. Bartalesi. Steps towards Enhancing the User Experience in Accessing Digital Libraries. In *Human Interface and the Management of Information. Information and Knowledge in Applications and Services – 16th Int. HCI Conference*, pages 555–566, Heraklion, Greece, 2014.

[Orlikowski and Yates, 1994] W. J. Orlikowski and J. Yates. Genre Repertoire: The Structuring of Communicative Practices in Organizations. *Administrative Science Quarterly*, (39):541–574, 1994.

[Rehm and Sasaki, 2016] G. Rehm and F. Sasaki. Digital Curation Technologies. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation (EAMT 2016)*, Riga, Latvia, May 2016. In print.

[Rehm, 2002] G. Rehm. Towards Automatic Web Genre Identification – A Corpus-Based Approach in the Domain of Academia by Example of the Academic's Personal Homepage. In *Proceedings of the 35th Hawaii International Conference on System Sciences (HICSS-35)*, Big Island, Hawaii, January 2002.

[Rehm, 2005] G. Rehm. *Hypertextsorten: Definition – Struktur – Klassifikation*. PhD thesis, Institut für Germanistik, Angewandte Sprachwissenschaft und Computerlinguistik, Justus-Liebig-Universität Gießen, 2005.

[Rumelhart, 1975] D. E. Rumelhart. Notes on a Schema for Stories. In Daniel G. Bobrow and Allan Collins, editors, *Representation and Understanding – Studies in Cognitive Science*, pages 211–236. Academic Press, New York, San Francisco, London, 1975.

[Sasaki *et al.*, 2015] F. Sasaki, T. Gornostay, M. Dojchinovski, M. Osella, E. Mannens, G. Stoitsis, P. Richie, T. Declerck, and K. Koidl. Introducing FREME: Deploying Linguistic Linked Data. In *Proc. of 4th Multilingual Semantic Web Workshop*, 2015.

[Y.-C. Li, 2012] W.-P. Su Y.-C. Li, A. Wee-Chung Liew. The Digital Museum: Challenges and Solution. *Information Science and Digital Content Technology*, pages 646–649, 2012.