The Structure of Social Documents

Susan P. Williams	Julian Mosen	Petra Schubert
University of Koblenz-Landau	University of Koblenz-Landau	University of Koblenz-Landau
Institute for Information	Institute for Information	Institute for Information
Systems Research	Systems Research	Systems Research
Faculty of Computer Science	Faculty of Computer Science	Faculty of Computer Science
Germany	Germany	Germany
williams@uni-koblenz.de	julianmosen@uni-koblenz.de	schubert@uni-koblenz.de

Abstract

Enterprise collaboration platforms are large scale, highly integrated information infrastructures that enable many hundreds of employees to work collaboratively and share information. In this paper, we lay the theoretical and analytical foundations for the use of social documents as digital traces of collaborative activity in enterprise collaboration platforms. Through a review of related research and an empirical analysis of social documents, we identify key concepts and structures, providing the foundation for the Social Document Ontology (SocDOnt). SocDOnt expresses the generic structure of social documents and extends previous work in two important ways. At the micro-level a social document is defined as a composition of an intellectual entity enhanced by both intellectual and simple components and at the macro-level a collection is defined as an aggregation of social documents. These analytical constructs enable a more nuanced and granular analysis of social documents to understand collaborative activity in enterprise collaboration platforms.

1. Introduction and motivation

Enterprise collaboration platforms are complex, large-scale information infrastructures comprising an ecosystem of highly integrated tools and functionality to support collaborative work and information sharing in organizations [23, 38]. In addition to their large-scale, integrated nature, the key difference between these platforms and previous collaboration systems is the native integration of enterprise social software (ESS) such as wikis, blogs, social profiles, activities, likes, tags etc. [34, 38]. ESS provides increased functionality for cooperative work and activity awareness, enabling employees to share, subscribe to, or follow information and people, and comment, tag or recommend the content created by other users.

Typically implemented in large organizations, enterprise collaboration platforms, (e.g. IBM Connections, Jive) are rich in ESS functionality and have become the de facto platform for the digital workplace. Used by organizations to span multiple global regions, business divisions and workgroups they support the collaboration, communication, coordination, content and knowledge sharing activities of many hundreds, often many thousands of employees and business partners, who are widely dispersed in both space and time [42].

The study of how people collaborate and work together has long been a focus of research in the fields of computer-supported cooperative work (CSCW) and information systems [17, 35]. However, a significant limitation of prior research is that it "appears to privilege particular forms of cooperative work" with many examples of "localist studies", restricted to particular settings and timeframes" [26:575–576]. To date, this is also the case in the context of enterprise collaboration systems, where empirical studies are often limited to cross-sectional studies of a single type of social software (e.g. blogs, wikis) [20, 34] or to a specific type of collaborative activity (e.g. knowledge sharing, project management) [1, 25]. This localist focus, often on single-site, small group interactions, is potentially problematic as today's organizations increasingly depend on *information infrastructures*, "large-scale, integrated and interconnected workplace information technologies", that are "typically stretched across space and time: [...] shaped and used across many different locales" and that "endure over long periods (decades rather than years)" [26]. Based on this, Monteiro et al. [26] argue that many research studies of collaborative work lack a large-scale, global view, and call for a broader perspective that accommodates "non-local constraints" and more "extended temporal scales" [26]. Enterprise collaboration platforms are a typical example of such an information infrastructure. They are highly complex software systems, spanning entire organizations and beyond, supporting many thousands of

group interactions and providing repositories of documented information that is often intended to be persistent and available to the organization and its employees over long periods of time. In addition, enterprise collaboration platforms are inherently malleable, they "begin life as empty shells" with no pre-existing content or inscribed work practices and "their meaning and value unfold over time and through users' interactions with the system" [28:581]. Users are free to choose which tools to use and how to use them to support their work. To date few studies have examined these large-scale, highly integrated enterprise collaboration platforms and the ways they are evolving and being shaped by users to support their organizational and collaborative work.

The research presented in this paper is part of a long-term program of empirical research that is addressing this limitation. For the past 10 years, through a university-industry research collaboration involving 38 industry partners we have been investigating the digital workplace and the use of large-scale enterprise collaboration platforms to support organizational work [41]. As part of our research we also host and manage a large-scale collaboration platform built around IBM Connections, currently one of the largest and most highly integrated commercially available enterprise collaboration platforms [15]. Our platform (UniConnect) is deployed as an academic collaboration system, enabling researchers from diverse, internationally distributed universities and research institutes to work together and organize collaborative research projects. The UniConnect platform currently hosts 35 universities and research institutions, has more than 3500 registered users and 1200 collaboration communities. In addition to being a fully deployed and operational system, UniConnect also provides our research team with a large-scale, information infrastructure for the observation, exploration, experimentation and evaluation of cooperative work, collaboration technologies and the digital workplace more widely [41]. The research program comprises a series of interrelated research streams, focusing on a range of topics including sociotechnical change and digital work, social collaboration analytics, social process mining and information integration in large systems.

In this paper, our focus is on the research stream directed towards obtaining a deeper theoretical and practical understanding of how collaborative work takes place within a collaboration platform. Our research investigates how employees are using the different affordances and functionality of the collaboration platform to develop new work practices and to organize their everyday work. To achieve this, we are examining the digital traces laid down when employees work together using a collaboration platform. Enterprise collaboration platforms support a wide range of work practices and provide many ways for people to work together to capture and share information, to coordinate team projects, and communicate and collaborate on joint work. All these activities leave digital traces in the collaboration system in the form of social documents [19]. *Social documents* include digital artefacts such as blog posts, wiki pages, forum topics, files, likes, tags and comments that are created as people collaborate on joint work. They are created "with the express intention of being interactive and collaborative" [19:48]. For example, an employee creates a blog post containing ideas for the development of a new product. This content is then extended by others who attach comments to the original content, add likes and tags and share it with other colleagues. These "attached" elements become important components of the original post and show how discussion and activity evolves around the original topic. By examining these social documents as *traces of collaborative activity* it is possible to gain insights into how employees are collaborating with each other to organize and coordinate work within the enterprise collaboration platform [13, 28]. However, a significant research challenge is gaining access to these document-mediated interactions and the methods to interpret them in ways that meaningfully identify collaborative activity. Enterprise collaboration platforms are large, complex systems containing a wide range of different collaboration and awareness tools (e.g. wikis, blogs, forums, tasks, activity streams, tags, likes) and document types (e.g. blog pages, wiki entries, comments, files) with diverse methods for interacting and using these highly integrated systems [36]. To analyze this collaborative activity first requires an understanding of the semantic structure underlying these complex artefacts and interactions. This paper addresses this requirement through a comprehensive investigation to identify, describe and represent the structure of social documents contained within an enterprise collaboration platform.

Our aim is to identify and clarify the structure of social documents and lay the theoretical and analytical foundations for using social documents as traces of collaborative activity in enterprise collaboration platforms. Specifically, our objectives are to analyze the semantic structure of the social documents being generated and derive a generic model to describe their structure at multiple levels from individual items (for example a single wiki page or comment) to complex collections of heterogeneous items. Thus, providing the theoretical foundation for the study of social documents and a basis for the development of methods and tools to visualize and analyze them in empirical settings. In addition, the in-depth understanding of the structure of social documents will also contribute to

research into improving their long-term management [19].

The paper is organized as follows. In Section 2 we examine related work to identify key concepts and terminology to provide a theoretical foundation for our empirical work. In Section 3 we present the research design for the in-depth study of the structure of social documents in enterprise collaboration platforms. In Sections 4 and 5, we present and discuss the study findings, its theoretical and analytical contribution, and their implications for future work on the use of social documents to trace collaborative activity in enterprise collaboration platforms.

2.   Documents as traces of collaborative activity

The study of documents as traces of activity in organizations has a long history in fields of research such as Library and Information Science and Records and Archival Studies [7, 9, 10, 12, 24] and more recently the study of digital documents in technology-mediated systems has formed a central stream of research in the fields of Workplace Studies and CSCW. The study of digital documents to support collaborative activity can be approached from several different perspectives [9, 33]. Digital documents can be examined as *structured artefacts* with clearly defined information models and metadata; acting as carriers of organizational information that can be integrated and exchanged between people and between diverse systems and technologies [14]. Documents can also be examined as *evidential records of organizational activities* [10, 39, 43] to not only investigate the individual document but also its context and provenance; offering insights into how it is related to, and interacts with other documents, entities and organizational processes and routines [16, 31]. Digital documents also act as *boundary objects mediating interaction, communication and collaboration* between people and with different technologies [29, 31]. It is against this theoretical background of documents as structured, evidentiary artefacts supporting interaction and communication that we locate our studies of collaborative activity in enterprise collaboration platforms.

Research to examine documents and documentary practices has a long history in the form of ethnographic studies of work [18, 30]. These studies provide very detailed and significant insights into collaborative work at the local and individual/group level, however they have a number of limitations for our proposed work in that they are often participant-observation based studies conducted around a specific task and

document type (e.g. electronic patient records [40]) or collaboration tools and activity (e.g. wikis and knowledge management [22]) and in specific locations within relatively narrow timeframes [13]. Our research to investigate collaborative activity in large-scale distributed enterprise collaboration platforms requires us to examine collaborative activity at the micro-level of the individual documents and task as well as at larger scales across workgroups and the platform as a whole. Thus, enabling us to follow the collective work practices of potentially thousands of users as they use a diverse range of tools and functionality to support their collaborative work practices. In addition, these insights will provide a basis for understanding the ways work practices are inscribed and how social documents and collaboration platforms evolve over time. However, before we can analyze collaborative activity we need to identify and understand the structure and nature of social documents contained within the collaboration platform, the purpose of the study presented in this paper and where our attention now turns.

2.1   The structure of social documents

In Section 1 we briefly presented the concept of social documents and their analytical potential to provide insights into collaborative activity in enterprise collaboration platforms. In this section we examine the structure of social documents more closely and draw on related research to identify key characteristics and clarify terminology. Social documents have been examined in two distinct, but related research views. In the fields of Information Studies and CSCW the focus is largely a *practice view* of the artefacts of collaborative activity [44]. In the fields of Web Sciences and Semantic Web, there is primarily a *representational view* of document ontologies with the objective to enable interoperability, integration and exchange of social documents between different tools and systems [6]. In both views, common concepts have emerged to describe the structure and nature of social documents, which are conceived as *compound documents that develop over space and time*.

In the practice view, social documents are defined as assemblages of related components (or fragments) created by multiple users collaborating on joint work [19, 44]. They are created in tools such as wikis, blogs and forums to "mediate the coordination of a widely distributed group committed to work towards a common goal" [44:206].

Important in these definitions is the distinction between the initial content, defined as the *intellectual entity* (e.g. a blog post or the initial wiki page) and the related *components* (e.g. comments, annotations, tags,

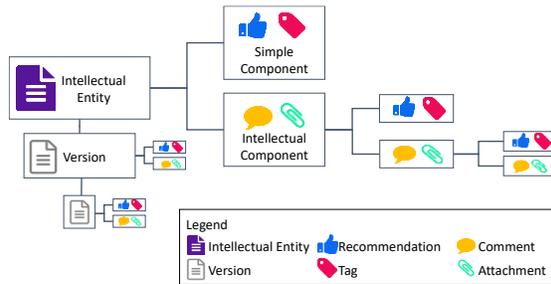links etc.) that are attached to the intellectual entity through subsequent collaborative actions.

Figure 1: Social document structure

Figure 1 illustrates the structure of a social document displaying the intellectual entity and its attached components.

Whilst providing fundamental concepts about the structure of social documents, previous work in this field has some limitations for our study of the structure of social documents in large-scale enterprise collaboration platforms. For example, Zacklad's work is primarily conceptual, and whilst providing a strong theoretical basis and argumentation for social documents as traces of collaborative activity it provides limited detail about their structure [44]. Further, whilst Hausmann and Williams [19] provide greater detail on social (business) documents, e.g. they identify different document types and provide illustrative examples of single types of social document structures, their work lacks a wider conceptual view to define the structure of multiple types of social documents and how they are combined and integrated within an enterprise collaboration platform.

These limitations are partially resolved by research from the representational view, which provides methods for "representing and navigating the content items … both within and across social websites" [6]. The goal of this work is primarily to develop representational mechanisms to interconnect people and objects in an interoperable and extensible way [5]. A potentially useful output of this work is the Semantically-Interlinked Online Communities (SIOC) ontology and its extensions, which "provides the main concepts and properties required to describe information from online communities (e.g., message boards, wikis, weblogs, etc.) on the Semantic Web" [3]. Our preliminary analysis of the SIOC Core Ontology Specification suggests that it offers a useful framework for defining the structure of social documents. However, our analysis also revealed some limitations of the SIOC ontology with regard to the definitions of social documents identified in the practice view of documents; in particular, the absence of the distinction between the intellectual entity and its attached components. In addition, there is to

date, limited empirical work that investigates the structure of social documents in everyday use in organizations.

3. Identifying the structure of social documents

In the following we present the research design and findings of an in-depth study of the structure of social documents in large-scale enterprise collaboration platforms. As outlined above, the main objective of the study is to identify and understand the generic structure of social documents. In this context, generic structure means the abstract description of all possible relationships between single content types, which are relevant for each other by contributing content or meaning and thus, must be considered as an aggregation to obtain the complete meaning of a social document. Understanding these structures and identifying aggregations of social documents is important for both research and practice. As per the main theme of this paper, identification of aggregations of social documents provides a basis for analyzing how people are collaborating and contributing to specific work tasks, thus adding to emerging research endeavors in the area of social collaboration analytics [38]. In a more practical setting, knowledge of the structure of social documents can be used to support work in the areas of records, archiving, legal discovery and regulatory compliance; where all parts of a document must be kept together and managed as a history and evidence of a specific matter or event. For example, in a legal discovery request all the comments and actions relating to a specific intellectual entity must be kept together for review purposes. Whilst previous research on social documents [19] is based on the investigation and comparison of one or only a few specific content types (e.g. wiki pages and blog posts), our research examines social documents within integrated enterprise collaboration platforms, covering the full range of collaboration features and types of social content. Thus, the intended outcome of our research is a comprehensive model that represents the generic structure of social content types.

3.1 Research design

The research design for the study comprises two interrelated phases of research as outlined below.

Research Phase 1: Social content analysis. In the first phase of work our focus is on the empirical analysis of social content. We investigate the implementation of social documents in an integrated collaboration

platform (IBM Connections). For this purpose, we use the UniConnect platform, described in Section 1. UniConnect is a fully functional, operational system, enabling us to conduct an in-depth examination of social documents and their structure. We do this from two perspectives: the user view and the technical system view for reasons explained below.

Social content analysis: user view. In the first step towards deriving the generic structure of social documents, we investigated their implementation in UniConnect from the viewpoint of the platform's user interface. Our objective, guided by the first three dimensions of the Social Collaboration Analytics Framework [38], was to identify i) *scope*: where social content can be created, ii) which *social content types* and iii) which *social components* can be created by a user. From the analysis, we identified three different areas of scope: the entire platform, a defined group workspace and a personal user workspace. In the case of UniConnect, six basic social content types were identified: files, forum posts, microblog posts, tasks, blog posts and wiki articles. Most of these types (but not all) can be enriched by four social components: attachments, comments, likes and tags.

Social content analysis: technical system view. In order to gain a deeper understanding of these elements we also conducted an in-depth analysis of the integrated enterprise collaboration platform to examine how social content and its structures are stored technically. Our aims were to find out which components of a social document can be identified within which type of data source and where the connections of single components are stored. Guided by the classification of data sources for Social Collaboration Analytics [38], we analyzed both user-generated *content data* and *organizational data*. While the content data stores the social content, the investigation of organizational data was necessary to identify the scope (platform, group workspace or user workspace) in which the social content is created. As the platform under analysis is based on proprietary, closed source software and does not provide any technical documentation on its architecture, it was necessary to perform a reverse engineering to extract the system's database schema. The advantage of this is that we looked directly at the system structure itself and derived an ER diagram for each database that stores content or organizational data.

Research Phase 2: Modelling social documents structures and harmonization. Following the analysis of social documents from the user's perspective and the technical point of view, we began modelling their structure for the specific case of UniConnect. Based on our findings from analyzing the databases, we transformed the entity relationship diagrams into a UML class diagram, describing all possible relations between the different content types and their components. Through an iterative process of evaluating, refining and generalizing the UML class diagram by a core team of five researchers we derived a first draft of a generic class diagram that describes the structure of social documents, independent of their underlying platform or system. The findings were also reviewed against other enterprise collaboration systems (including Alfresco, Sharepoint and Atlassian Confluence) in order to evaluate their completeness in the context of currently available systems.

In the final step, we synthesized the findings from the modelling and analysis to develop the *Social Document Ontology* (SocDOnt). The details of SocDOnt are presented and discussed in the following section. As part of the process of transforming the former UML class diagram into an ontology, and to ensure harmonization with existing work, we analyzed related ontologies to compare our findings with existing concepts and terminologies. We primarily focused on the *Socially Interlinked Online Communities Ontology* (SIOC) [2], which provides a comprehensive data model for machine-readable processing and interoperability of content from online communities [32]. In addition to its core concepts, SIOC has been extended by single modules, such as the *SIOC Types Module* (SIOCT), which contains the description of further content types for online communities [4]. In our extended ontology, following the recommendation of Passant et al. [32], we make use of concepts, properties and attributes from existing ontologies wherever possible "to avoid reinventing new classes and properties, and to benefit from past work (…) in terms of ontology engineering" [32:184]. This enabled us to align SocDOnt with existing ontologies, to harmonize our terminology and identify new concepts and requirements to describe the structure of social documents in enterprise collaboration platforms.

4. Findings and implications for social document analysis

The *Social Document Ontology (SocDOnt)* provides a comprehensive model and a terminology for the description of social document structures. A complete overview of SocDOnt, represented as a UML class diagram, is shown in Figure 2. SocDOnt makes use of concepts from existing ontologies, such as the SIOC ontology, its extension SIOCT and the *Task Management Ontology* (TMO) [8] for describing the concept of tasks, and introduces new concepts, necessary for a more detailed description of social documents within enterprise collaboration platforms.
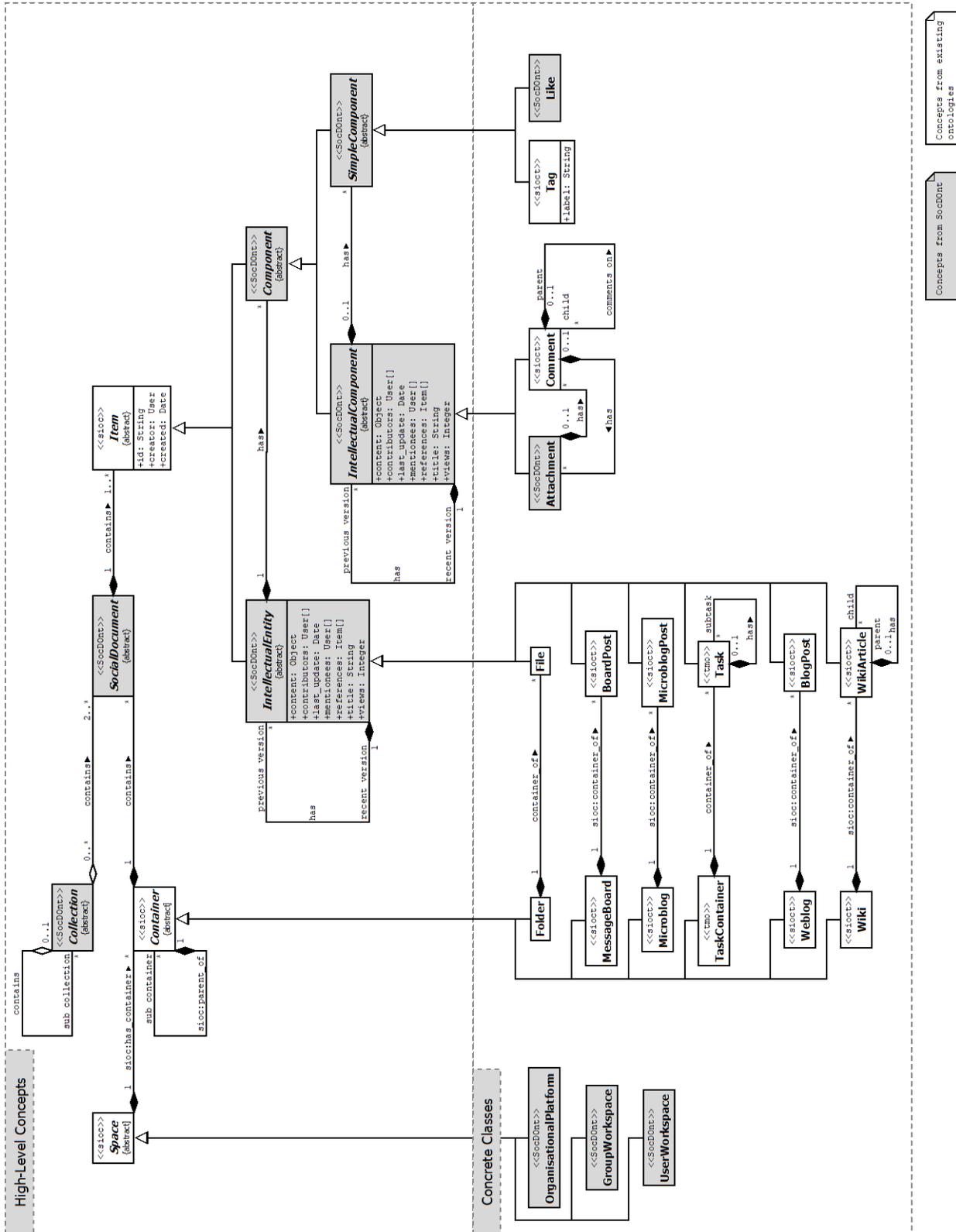
Figure 2. UML-based representation of the Social Document Ontology (SocDOnt)

The prefixes of the class names and associations in Figure 2 indicate the originating ontology the concepts are derived from. The prefixes *«sioc»* and *«sioct»* indicate that a concept is described by SIOC or SIOCT and the concepts native to SocDOnt are labelled with the prefix *«SocDOnt»*. The *top section* of the diagram shows high-level concepts (abstract classes), which describe the generic structure of social documents, the *lower section* contains concrete classes, which represent exemplary types identified in the UniConnect platform and typically found in all the systems we analyzed.

It is important to stress that most associations in the UML class diagram are modelled as *compositions* (associations with filled diamond shape) instead of *aggregations* (hollow diamond shape) as the characteristics of a composition are better suited for describing the nature of social documents and their components. That is, the subordinated components only have meaning in the context of the superordinate item. For example, when a social document or a component is deleted, all subordinated objects should be deleted as well (e.g. if a blog post is deleted, its comments and recommendations lose their meaning and should be deleted). Similarly, when a social document is archived, it is important for evidentiary and compliance purposes, that all components of the document are archived together. For example, all the comments related to a blog post should be archived along with the originating post. In the following, we describe and explain the key elements of SocDOnt, working from right to left on Figure 2.

Social Documents and Items. The core of the ontology is built by the concept of a *social document*. In the context of SocDOnt a social document is an abstract object, describing a composition of tightly connected social content *items* that are not separable, partly or at all, without the loss of meaning. An *item* is a single piece of social content, e.g. a blog post or comment, and refers to the correspondingly named concept from the SIOC ontology. A good example for the inseparable composition of items is a forum thread, which is composed of multiple items, such as an initial post, related comments, tags and attachments, which only make sense within the context of the related forum thread. The initial item of a social document is its *intellectual entity*. A social document has exactly one intellectual entity, which can occur in different forms. Within UniConnect, we identified six types of intellectual entities: files, board posts, microblog posts, tasks, blog posts and wiki articles. With the exception of uploaded files, each type of intellectual entity is described by a correspondingly named concept in SIOC, SIOCT or TMO. All types of intellectual entity have common attributes, which are inherited from one of their parent

classes (*IntellectualEntity* and *Item*). These attributes contain a unique identification number (e.g. for referencing a component via hyperlinks), a creator representing the person who created it, timestamps indicating when it was created and updated, some form of intellectual content (e.g. the text in a blog post), a list of people who are mentioned in the content (e.g. via @mentions), a list of people who contributed the content, a list of items, which are referenced within the content (e.g. via hyperlinks), a title (e.g. blog post title) and the number of views (e.g. page views from a web browser). Importantly, the social document is not static but dynamic; it can be changed and edited over time. Many platforms offer a versioning feature that automatically creates a new *version* of an intellectual entity if it is edited or updated. This leads to intellectual entities having at least one current version and multiple previous versions. Some types of intellectual entities can have reflexive associations indicating a parent-child relationship, which is very typical for wiki articles (parent and child wiki articles) or tasks (tasks and subtasks).

In addition to the intellectual entity, a social document can have further items, which are described as social document *components*. In contrast to intellectual entities, components cannot exist on their own but must be associated with exactly one item of a social document, either an intellectual entity or another component. The SocDOnt introduces two types of components: *intellectual components* and *simple components*. Simple components typically occur in the form of tags and likes (recommendations) and can be associated to an intellectual entity or an intellectual component (attachment and comment). Due to their lack of intellectual content, tags and likes can never be associated with each other or another instance of themselves (tagging a tag, liking a like or tagging a like does not make sense; liking a tag might be possible but does not occur in our system). In the context of SocDOnt, reused tags having the same name are modelled as multiple instances. In contrast to modelling equally named tags as a single instance, multiple instances allow the storage of additional meta data within the inherited attributes from the parent class *Item*, such as the user (creator) who attached the tag to an item and the corresponding date (created). Intellectual components (attachments and comments) differ from simple components by reason of their intellectual content, which could be, for example, the content of an attached file or information in a comment. Intellectual components are similar to intellectual entities, have the same attributes and might even be tracked with versioning features. In contrast to simple components, intellectual components can be associated with each other because of their intellectual content (e.g. commenting on an attachment, attaching

an attachment to a comment). While reflexive associations of attachments are unusual (but theoretically possible), such kinds of associations are very typical for comments (i.e. commenting on a comment). The characterization of social documents as i) comprising an intellectual entity and associated components and ii) distinguishing between intellectual components and simple components draws from concepts identified in the practice-view of documents outlined earlier (e.g. [19, 44]). This represents an important theoretical and practical extension to work in, for example, the SIOC ontology, where no distinction is made and all items are equal.

Containers. Enterprise collaboration platforms contain different applications (social features or modules) that can be used for creating and storing social content. While components can be created and attached to an item within each (or most) applications, each type of intellectual entity can only be created within one dedicated type of application (e.g. a blog post can only be created within its application "weblog"). Social documents and items created in the same application are stored in *containers*. The concept of a container is defined by the SIOC ontology and describes a high-level concept for grouping items that are created and stored by the same application [2]. Containers are either created automatically in the application or manually by the user. In the case of UniConnect, we identified six types of containers for social documents: file folders, message boards (forums), microblogs, task containers (activities), weblogs and wikis. The existence of a container is mandatory for the existence of a social document (e.g. a blog post is always part of a weblog and cannot exist on its own). Like the reflexive associations of intellectual entities (tasks and wiki articles), containers can be nested as well (e.g. folders and subfolders).

Spaces. In enterprise collaboration platforms, containers and social documents are created and stored in a specific (work)*space*. The SIOC ontology describes a space as a "place where data resides" and a "location for a set of Container(s) of content Item(s)" [4:30]. SocDOnt makes use of the space concept and introduces three subclasses: O*rganizational platform*, *group workspace* and *user workspace*. The organizational platform space describes the entire space of in enterprise collaboration platform and includes all social documents that are not stored in workspaces or user spaces. Group workspaces can be public or only allow restricted access for a defined group of users and user workspaces describe the personal workspace of individuals.

Collections. As described earlier, the intellectual entity of a social document can have reflexive associations within its container (e.g. a wiki article with sub-

pages). Additionally, intellectual entities and intellectual components can contain references to other intellectual items, which can be part of another social document (e.g. a blog post quoting or linking a wiki article). In contrast to reflexive associations, references are not limited to items within the same container but can occur across containers and spaces. Both, reflexive associations of intellectual entities and references between intellectual items, result in an association of two or more social documents. SocDOnt describes these types of connected social documents as *collections* and defines them as an **aggregation of social documents that are associated with each other**. The term collection and its concept in SocDOnt are guided by the correspondingly named concept from the *Dublin Core Schema* (DC), which describes it as an aggregation of items [11]. It is important to note the use of the term "aggregation" here, referring to the fact that a single social document can be deleted from a collection and the collection still has meaning. Further definitions from the field of archival science describe that collections "may be grouped in hierarchical structures" [21]. In SocDOnt, these hierarchical structures of collections correspond to reflexive associations of intellectual entities, leading to nested collections, for example, a wiki itself represents a top collection, which contains many wiki articles. If a wiki article has subpages (child articles), a wiki article becomes a sub collection of the wiki (top collection). The same phenomena can be observed for tasks, which can be composed of subtasks. An important theoretical and analytical distinction is drawn between a collection and a container. A container is created intentionally by a user as a place to put content (e.g. in the case of a folder), whereas a collection is formed over time, as social documents are created, linked and extended by different users. Further, a collection may be spread across different containers (e.g. when a user attaches a file [container: files] to a blog post [container: weblog]) or across spaces (e.g. sharing and referencing a file from a personal workspace with a group workspace). In contrast to the concept of a container, which can be identified as a concrete instance within an enterprise collaboration platform, collections are non-physical aggregations of social documents based on their semantic connections. The fact that collections extend beyond the boundaries of containers and spaces, means they offer new possibilities for capturing the *macro-level structure* of social documents that are less storage-centric and more holistic and practice-oriented. Collections provide analytical potential as traces of collaborative activity, extending the scope of investigations to examine collaboration across space and over time.

5.   Summary and conclusions

The aim of this paper was to lay the theoretical and analytical foundations for the use of social documents as traces of collaborative activity in enterprise collaboration platforms. To achieve this requires a detailed understanding of the semantic structure underlying social documents and their interactions within a collaboration platform. Through a review of prior research and an empirical analysis of social documents in an operational collaboration platform we identified key concepts and structures. We used these findings to develop the SocDOnt ontology to represent the generic structure of social documents.

The comparison of SocDOnt with existing ontologies ensured the harmonization of concepts and terms wherever possible, however it also revealed a number of limitations in their application in the context of a collaboration platform. Whilst providing a useful foundation, the existing SIOC ontology was developed for online communities and focuses on public, independent spaces, containing only one or a small number of social software features. It was not developed to model social content within enterprise collaboration platforms consisting of many integrated applications. Our research showed that existing ontologies, such as SIOC, provide a suitable foundation for describing some aspects of social document structures, but do not provide sufficient capabilities for their generic description on a *micro* level (intellectual entity and components) or a *macro* level (social documents and collections) within enterprise collaboration platforms.

More specifically, at the *micro level*, drawing from work in the field of Library and Archival Studies [19, 21, 44] we accommodate the concept of the social document as a composition of an intellectual entity and related components, the intellectual entity being the core of the social document and having primacy. Components, representing subsequent collaborative activity may then be related to the intellectual entity. Further, we distinguish between intellectual and simple components, enabling us to understand collaborative activity in more granular detail. In the SIOC ontology, this distinction between item types is not made and all items have the same valency.

At the *macro level* the generic description shows social documents as created and stored in containers and potentially being part of multiple collections. A collection is an aggregation of multiple social documents, which are interconnected. Social documents that are part of the same collection can be spread across different containers and spaces; this representation is not included in the SIOC ontology.

The *intellectual entity* and *collection* are both important analytical constructs for social document ana-

lysis. They enable us to now analyze how collaborative activity takes place by identifying, at both a micro and macro level, where work begins (intellectual entity) and then tracing how it evolves through collaborative activity to add components, link to other social documents and form collections. SocDOnt provides the necessary basis for social document analysis and for tracing collaborative activity over both *space and time* and it is *extensible*. If new types of containers and items are developed in the future, they can be included in the respective areas of SocDOnt.

The research and development of SocDOnt have provided a necessary foundation and enabled us to continue our research work through the development of methods for visualizing [27] and analyzing [37] social content to trace collaborative activity in enterprise collaboration systems.

6.   References

[1] Bean, L., and D.D. Hott, "Wiki: A speedy new tool to manage projects", *Journal of Corporate Accounting & Finance 16*(5), 2005, pp. 3–8.

[2] Bojars, U., and J.G. Breslin, "SIOC Core Ontology Specification Revision 1.26", *W3C Member Submission*, 2007. https://www.w3.org/Submission/2007/SUBM-sioc-spec-20070612/

[3] Bojars, U., and J.G. Breslin, "SIOC Core Ontology Specification Revision 1.36", 2018. http://rdfs.org/sioc/spec/

[4] Bojars, U., J.G. Breslin, V. Peristeras, G. Tummarello, and S. Decker, "Interlinking the social web with semantics", *IEEE Intelligent Systems 23*(3), 2008, pp. 29–40.

[5] Breslin, J.G., and S. Decker, "The Future of Social Networks on the Internet: The Need for Semantics", *IEEE Internet Computing 11*, 2007, pp. 86–90.

[6] Breslin, J.G., A. Passant, and S. Decker, *The Social Semantic Web*, Springer, Berlin Heidelberg, 2009.

[7] Briet, S., *What is documentation? translated and edited by Ronald E. Day and Laurent Martinet*, Scarecrow Press, Lanham, MD, 2006.

[8] Brunzel, M., and O. Grebner, "TMO - Task Management Ontology", 2008. http://oscaf.sourceforge.net/tmo.html

[9] Buckland, M., "What is a 'document'?", *Journal of the American Society for Information Science 48*(9), 1997, pp. 804–809.

[10] Choksy, C.E.B., *Domesticating information: Managing documents inside the organization*, Scarecrow Press, Lanham, MD, 2006.

[11] Dublin Core Collection Description Task Group, "Dublin Core Collection Description Application Profile", *The Dublin Core Metadata Initiative (DCMI)*, 2007. http://www.dublincore.org/specifications/dublin-core/collection-description/collection-application-profile/2007-03-09/

[12] Frohmann, B., *Deflating information: from science studies to documentation*, University of Toronto Press, Toronto, 2004.

[13] Geiger, R.S., and D. Ribes, "Trace ethnography: Following coordination through documentary practices", *44th Hawaii International Conference on System Sciences*, (2011), 1–10.

[14] Glushko, R.J., and T. McGrath, *Document Engineering: Analyzing and Designing Documents for Business Informatics and Web Services*, The MIT Press, Cambridge, Mass, 2005.

[15] Gotta, M., N. Drakos, and M. Jeffrey, *The Future of Social Software in the Workplace*, Gartner Research Report ID: G00293232, 2017.

[16] Greenhalgh, T., "Role of routines in collaborative work in healthcare organisations", *British Medical Journal 337*, 1998, pp. 1269–1271.

[17] Greif, I., *Computer-supported cooperative work: a book of readings*, Morgan Kaufmann Publishers Inc., San Mateo, CA, 1988.

[18] Harper, R., *Inside the IMF: An ethnography of documents, technology, and organizational action*, Academic Press, San Diego, 1998.

[19] Hausmann, V., and S.P. Williams, "Issues for the long-term management of Social Business Documents", *International Journal of Information Systems and Project Management 4*(3), 2016, pp. 45–61.

[20] Holtzblatt, L.J., L.E. Damianos, and D. Weiss, "Factors Impeding Wiki Use in the Enterprise: A Case Study", *28th ACM Conference on Human Factors in Computing Systems*, (2010).

[21] InterPARESTrust, "InterPARES Trust Terminology - Collection", 2019. https://interparestrust.org/terminology/term/collection

[22] Jackson, P., and J. Klobas, "Deciding to use an enterprise wiki: the role of social institutions and scripts", *Knowledge Management Research & Practice 11*, 2013, pp. 323–333.

[23] Leonardi, P.M., M. Huysman, and C. Steinfield, "Enterprise Social Media: Definition, History, and Prospects for the Study of Social Technologies in Organizations", *Journal of Computer-Mediated Communication 19*(1), 2013, pp. 1–19.

[24] Levy, D.M., *Scrolling forward: making sense of documents in the digital age*, Arcade Publishing, New York, 2001.

[25] Millen, D.R., M.J. Muller, W. Geyer, E. Wilcox, and B. Brownholtz, "Patterns of media use in an activity-centric collaborative environment", *CHI 2005*, (2005), 879–888.

[26] Monteiro, E., N. Pollock, O. Hanseth, and R. Williams, "From Artefacts to Infrastructures", *Computer Supported Cooperative Work 22*(4–6), 2013, 575–607.

[27] Mosen, J., S.P. Williams, and P. Schubert, "Visualizing Social Documents as Traces of Collaborative Activity in Enterprise Collaboration Platforms", *53rd Hawaii International Conference on System Sciences*, (2020).

[28] Nitschke, C.S., and S.P. Williams, "Traces of design activity: the design of coordination mechanisms in the shaping of enterprise collaboration systems", *CENTERIS Conf. on ENTERprise Information Systems*, (2018).

[29] Olsen, B.I., N.W. Lund, G. Ellingsen, and G. Hartvigsen, "Document theory for the design of socio-technical systems", *Journal of Documentation 68*(1), 2012, pp. 100–126.

[30] Østerlund, C., "Documents in Place: Demarcating Places for Collaboration in Healthcare Settings", *Computer Supported Cooperative Work (CSCW) 17*(2–3), 2008, pp. 195–225.

[31] Østerlund, C., and K. Crowston, "Boundary-spanning documents in online communities", *Thirty Second International Conference on Information Systems*, (2011), 1–10.

[32] Passant, A., U. Bojars, J.G. Breslin, et al., "The SIOC Project: Semantically-Interlinked Online Communities, from Humans to Machines", *Coordination, Organizations, Institutions and Norms in Agent Systems V*, Springer Berlin Heidelberg (2010), 179–194.

[33] Pédauque, R., *Form, Sign and Medium, As Reformulated for Electronic Documents (Working Paper sic_00000594 2003)*, 2003.

[34] Richter, A., and K. Riemer, "Malleable end-user software", *Business and Information Systems Engineering 5*(3), 2013, pp. 195–197.

[35] Schmidt, K., and L. Bannon, "Constructing CSCW: The First Quarter Century", *Computer Supported Cooperative Work 22*, 2013, pp. 345–372.

[36] Schubert, P., and J.H. Glitsch, "Use Cases and Collaboration Scenarios: How employees use socially-enabled Enterprise Collaboration Systems (ECS)", *International Journal of Information Systems and Project Management 4*(2), 2016, pp. 41–62.

[37] Schubert, P., J. Mosen, and F. Schwade, "Metrics for Analyzing Social Documents to Understand Joint Work", *53rd Hawaii International Conference on System Sciences*, (2020).

[38] Schwade, F., and P. Schubert, "Social Collaboration Analytics for Enterprise Collaboration Systems: Providing Business Intelligence on Collaboration Activities", *50th Hawaii International Conference on System Sciences*, (2017), 401–410.

[39] Spinuzzi, C., *Tracing Genres through Organizations: A Sociocultural Approach to Information Design*, MIT Press, Cambridge, MA, 2003.

[40] Ventres, W., S. Kooienga, N. Vuckovic, R. Marlin, P. Nygren, and V. Stewart, "Physicians, Patients, and the Electronic Health Record: An Ethnographic Analysis", *Annals of Family Medicine 4*(2), 2006, pp. 124–131.

[41] Williams, S.P., and P. Schubert, "Connecting Industry: Building and Sustaining a Practice-based Research Community", *50th Hawaii International Conference on System Sciences*, (2017), 5400–5409.

[42] Williams, S.P., and P. Schubert, "Designs for the Digital Workplace", *Procedia Computer Science*, Elsevier B.V. (2018), 478–485.

[43] Yates, J., *Control through communication - The rise of system in American management*, Johns Hopkins University Press, Baltimore, 1989.

[44] Zacklad, M., "Documentarisation processes in documents for action (DofA): The status of annotations and associated cooperation technologies", *Computer Supported Cooperative Work 15*, 2006, pp. 205–228.