# On the Notion of Genre in Digital Preservation*

Fiorella Foscarini, Yunhyong Kim, Christopher A. Lee,
Alexander Mehler, Gillian Oliver, and Seamus Ross

**Abstract**

In this paper, we discuss the notion of genre as a basis for addressing the problem of context representation in digital preservation. We outline several reference points for the notion of genre. This includes a review of diplomatic principles that can support and enhance the power of genre as a key to capture information about context relations. Further, we discuss the impact of open genre models and open topic models in information retrieval and finally present a list of research questions concerning future research in automation of digital preservation.

**Keywords:** *digital preservation, genre, context modeling, diplomatics, information retrieval*

## 1 Setting the Scene

Digital preservation aims to ensure that digital objects of value to society – whether, for instance, as a foundation for future discoveries or as evidence of human activities – can be meaningfully reproduced over time, despite evolving representation mechanisms, rapidly advancing technologies, and continually emerging user expectations. Future users will encounter digital objects that have been subjected to a variety of preservation actions, ranging from substantial transformative migrations of the objects themselves to imitation of earlier computer environments through emulation. Regardless of which preservation strategies have been applied, the future users are likely to imagine that the characteristics and relationships of digital objects that they are accessing are significantly similar to the characteristics and relationships of the digital objects when they were created and first used. However, minor differences can "make a difference" (Bateson, 1972) in how digital objects are experienced, understood and valued.

In other words, digital preservation does not simply involve maintaining physical data carriers (i.e. storage media) or reproduction of bit streams from the carriers. Rather, as artifacts, digital objects (e.g., documents, pictures, videos, online games, social software) are complex aggregations of signs (Peirce, 1934) whose form and meaning vary depending on various levels of situational context (Barwise and Perry, 1983). Those who are taking digital preservation actions should be cognizant of the representation of information within the semiotic triad of syntax, semantics and pragmatics (Morris, 1938). It

*In: Proceedings of the Dagstuhl Seminar 10291 on *Automation in Digital Preservation*, July 18–23, 2010, Schloss Dagstuhl.



Dagstuhl Seminar Proceedings 10291
Automation in Digital Preservation
http://drops.dagstuhl.de/opus/volltexte/2010/2763

is widely recognized that "bit preservation" that focuses solely on the material aspects of digital objects (that is, on the "digital substrate" of their form, Hjelmslev 1969) is not sufficient, because it does not address syntax and semantics at the full range of levels at which they are inherent in, and significant to the understanding of, digital objects. Moreover, bit preservation does not reflect pragmatics; a concept which encapsulates the social as well as situational patterns that influence the way corresponding discourse communities make use of objects. In order for users to make sense and meaningful use of artifacts at moments of "reactivation" in the future, it will often be important to express information about the artifacts' original contexts of creation and use (Lee, 2010). Pragmatic context is never exhaustively conveyed by the artifacts themselves. Diplomatics provides a set of conceptual and methodological tools for inferring aspects of functional context from the form and/or structure of the documents themselves. However, when the objects do not present a fixed structure or stable content – often the case in today's digital environment – it can be vital to capture or create additional information (contextual units) that can be conveyed to users of the objects if those users are to understand them.

All material things deteriorate over time, and loss is a fact of life. We can only preserve traces of objects' existence and evidence of their value throughout their life cycle, so that future generations may understand something of their legacy even after material elements have been lost. So the fundamental question is: "how do we determine value?" While this is not an easy question, part of the answer lies in determining how objects support core human activities (e.g. research, communication, administration, organization, transaction). To this end, we have been drawing on the expertise of archivist and library science specialists who have long been educated in the arts of appraising, selecting, record-keeping, curating and storing material for the sake of documenting, evidencing, and supporting human activity within operational and cultural communal, organizational and business settings.

The digital preservation community's scope of interest and responsibility is not confined to the closed environment of an organization. The community's objectives extend to larger networks of associated entities, and further, to the Internet. It seems reasonable, then, for us to consider the nature of human activities (e.g. information retrieval, blogging) that have emerged within this wider context of use and to question whether current preservation practices are adequate to document traces of these activities.

Within the literature about digital collections, we have seen an increasing focus on information context and use, reflecting the (re)emerging awareness of signifiers of information value that go beyond simple subject matter. Signifiers can reflect the functional aspects of the material (e.g. to record processes, to advertise products, to describe research profiles, to convert people to your way of thinking), attitudes of the actors involved (e.g. polarity or opinions or sentiments), literary style (e.g. author attribution), and relevance criteria (e.g. currency, document novelty, depth and scope – cf. (Cerviño Beresi et al., 2010)). Some of these aspects have been presented elsewhere as aspects of genre. The use of genre as a structural aspect of text to aid the automated extraction of metadata for the management of digital material has been examined previously (Kim and Ross, 2006). In this paper, we discuss the notion of genre as a basis for addressing the problem of context representation. We outline several reference points for the notion of genre, including

a review of diplomatic principles that can support and enhance the power of genre as a key to understand and capture information about context relations.

## 2 What is Genre

There is, in general, a lack of consensus in the literature relating to the definition of genre (Kim and Ross, 2007c,b; Santini et al., 2010), and this is a reflection of the complexity of the domain and the diversity of ways in which the concept of genre can be effectively deployed.

In this paper, we adopt a broad definition of genre as a socially recognized pattern of communication that conforms to established expectations. Examples of genres will encompass the whole gamut of communicative acts including speech utterances, text messages, publications, databases, and images. Genres reflect the traditions of their production, creation and use as artifacts, and should not be confined to the notion of being just another text type (Spinuzzi, 2003). They "[...] represent the development and stabilization of worldviews, including the values, ethics, and other humanistic concerns implied in them" (Spinuzzi, 2003, p. 41). This suggests that genres, and the classification of digital objects into different (and often multiple) genre classes, can contribute many forms of contextual information that would otherwise be missing.

Genre as a socially recognized pattern of communication fits well into other notions, including genre systems (Yoshioka et al., 2001; Østerlund, 2007), genre as social action (Miller, 1984), genre ecologies (Bawarshi, 2001; Spinuzzi, 2003), and genres as evolving clusters of family resemblances (Wittgenstein, 1953; Kim and Ross, 2007a). These concepts are significant because they reflect the embeddedness of documents or digital objects within communicative social environments.

Yates and Orlikowski define genres as "typified communicative actions invoked in response to a recurrent situation" (Yates and Orlikowski, 1992, p. 301). When the "recurrent situation" involves facts that are juridically relevant, any manifestations of communicative actions – especially written ones – will tend to comply with specific rules of representation, as dictated by the concerned juridical-administrative context. These are the situations in which the concepts and methods of diplomatics traditionally apply.

## 3 A Representation Model of Genre: Diplomatics

Investigations of the nature of documents, their relationship to processes which led to their creation and in which they were used, and the place of genre in handling (e.g. classifying them) and understanding them requires a theoretical and methodological framework. Diplomatics provides such a framework to underpin analysis of the characteristics of the *written evidence* of the facts and acts having a juridical nature. It is based on the assumption that there is a direct relationship between the structural features, or 'elements of form', of documents and the administrative procedures or business activities originating them.

Diplomatics emerged in the second half of the 17th century as an analytical method for determining the *authenticity* of medieval charters issued by sovereign authorities, for the ultimate purpose of ascertaining the truthfulness of the facts represented in them.[1] Jean Mabillon, author of *De re diplomatica* (1681), established the general principles of diplomatics, including "a corpus of references which would enable the testing of a document and each of its elements, one after the other" (Guyotjeannin, 1996, p. 415). During the 19th century, in most European countries, diplomatics became part of the body of knowledge of the archival discipline and has continued to evolve as a tool for the retrospective understanding of historical sources (Boyle, 1976). With the emergence of electronic records and the challenge of identifying, making sense of, and preserving the circumstances of their creation and use, archivists have rediscovered the fundamental contribution that diplomatics can provide for learning about documentary, administrative, and juridical contexts. Luciana Duranti, the archival scholar who broadened the applicability of diplomatic concepts and methods to contemporary bureaucracies, defines diplomatics as "the discipline which studies the genesis, forms, and transmission of the documents, and their relationship with the facts represented in them and with their creator, in order to identify, evaluate, and communicate their true nature" (Duranti, 1998, p. 45).

The study of the *genesis* of documents attempts to establish the details and steps involved in the activities that generate documentary evidence. This involves analyzing on the one hand, the mechanisms of decision-making (i.e., the *moment of action*) and on the other, the correspondent procedures governing the formation of documents (i.e., the *moment of documentation*). Diplomatics considers the genesis of documents as an elaboration of routines. Diplomatics identifies the typical components of the procedures guiding both action and documentation, which are evident in the formal elements of documents, based on the assumption that every procedure, or transaction, tends to be structured. Independently of its context, author and purpose, the "ideal structure" of each single procedure would comprise two or more of the following steps:

1. *initiative*;
2. *inquiry*;
3. *consultation*;
4. *deliberation*;
5. *deliberation control*;
6. *execution*.

[1] *Diplomatic authenticity* does not coincide with *legal authenticity*, although both can lead to an attribution of *historical authenticity* in a judicial dispute. "Legally authentic documents are those which bear witness on their own because of the intervention, during or after their creation, of a public authority guaranteeing their genuineness. Diplomatically authentic documents are those which were written according to the practice of the time and place indicated in the text, and signed with the name of the person competent to create them. Historically authentic documents are those which attest to events that actually took place or to information that is true" (Duranti, 1998, pp. 45-46).

In the first three steps, all documents created have an interlocutory nature with respect to the transaction as a whole. In the subsequent three steps, the focus of each action is the preparation, completion and perfecting of the documents embodying the transaction. An understanding of these procedural steps and their relationships with the formation of relevant documentary outputs is essential to archivists, as the form and the status of transmission (i.e., original, copy, or draft) of the documents they encounter results from the status of development of the procedures generating them.

Diplomatics divides possible procedures into four categories, based on their general purpose:

1. *organizational procedures* (aiming at the establishment, modification, or extinction of organizational structures and internal rules);
2. *instrumental procedures* (relevant to the expression of opinions or advice);
3. *executive procedures* (allowing for the regular transaction of affairs according to existing norms); and
4. *constitutive procedures* (those which create, extinguish, or modify the exercise of power; and which comprise the following subcategories: *procedures of concession*; *procedures of limitation*; and *procedures of authorization*).

Although today's bureaucratic environments are more complex than those studied by the diplomatists dealing with medieval records, the above categorization retains its explanatory and expository validity. In any juridical systems, human endeavors always present an organizational, an instrumental, and an executive or a constitutive nature. One difference is that the four types of procedures can be found today at many levels rather than at one level only. Also in contrast to the medieval context, when each given documentary form was likely to be the result of one specific procedure, in the modern context, procedures which have different purposes often create the same documentary forms, and procedures having the same purpose often produce different documentary forms. However, this just reinforces the usefulness of 'mapping' documentary products against the functions and activities of their creators by reconstructing and examining the procedures of document creation and the ways in which all these actions manifest themselves in documentary forms.

Diplomatics as a method of inquiry is based on the observation that "the *form* of a document reveals and perpetuates the function it serves" (Duranti, 1998, p. 133). Documentary forms are both physical and intellectual. The physical form refers to the external make-up of the document and includes medium, script, language, special signs, seals, and annotations – known as *extrinsic elements of form*. Every moment in history, every country, every profession or social group is characterized by the use of certain formulae, bureaucratic or literary styles, specialized languages, and so on. The 'critique of the false' as practiced by medieval and modern diplomatists relies upon the knowledge of documentary practices derived mainly from the observation of the extrinsic elements of form. For contemporary materials, there is no comprehensive catalogue of document typologies (or genres) (Barbiche, 1996).

The intellectual form involves the *intrinsic elements of form* and refers to the internal articulation of the document, that is, the mode of presentation of its content, "the parts of the discourse [that are] necessary to tell who does what for whom, why, where, when and how" (Guyotjeannin, 1996, p. 417).[2] Diplomatics assumes that all documents present an "ideal analytical sub-structure" (Pratesi, 1962, p. 62) where the intrinsic elements of form may appear and which involves three sections, each having a specific purpose.

1. The first section is called *protocol* and contains the administrative context of the action, i.e., an indication of the persons involved (entitling, superscription, inscription), time and place (chronological and topical date), subject and other optional elements (invocation, salutation, appreciation, etc.).

2. The second section is the *text*, the central part of the document, where one finds the manifestation of the will of the author. It usually starts with a preamble, which expresses the ideal (ethical or juridical) motivation for the action. In official documents, this part is often followed by a notification (i.e., the publication of the purport of the document). The substance of the text is introduced by the exposition of the circumstances generating the act/document and is followed by the disposition (i.e., the expression of the will, usually through a verb able to communicate the nature of the action and function of the document, such as, authorize, promulgate, request, etc.). Final clauses might end the text.

3. The third section is the *eschatocol*, which mainly consists of the attestation (i.e., the subscription of those who took part in the issuing of the documents and of any witnesses) and may include the date as well as other formulas and clauses (corroboration, complimentary clause, qualification of signature, secretarial notes, etc.).

Diplomatics offers a sophisticated understanding of the form of a document as "the complex of the rules of representation used to convey a message" (Duranti, 1998, p. 41). Through a process of decontextualization, early diplomatists could devise one *ideal documentary form* which includes all the elements examined above. Diplomatic criticism consists in analyzing the variations and presence or absence of any of those elements in actual documentary forms in order to reveal the functions of documents manifesting the forms and to establish documents' identity and integrity (Duranti, 1998, p. 134). Diplomatic concepts and methods may be employed in contemporary information systems to facilitate the 'automatic' extraction of functional (i.e., contextual) knowledge from textual objects, whether on paper or electronic.

In particular, "diplomatics comprises a body of concepts and principles that provide a strong conceptual model of an *authentic* record" (MacNeil, 2004, p. 223). The main

[2]This diplomatic notion resembles the six dimensions of communicative interaction identified by Yates and Orlikowski, who write: "Genre systems are organizing structures [...] providing expectations about the purpose, content, form, participants, time, and place of communicative interaction. In other words, [...] the why, what, how, who/m, when and where" (Yates and Orlikowski, 2002, p. 16). Genre theory does not elaborate on the formal elements embodying those dimensions, while diplomatics offers an elaborate characterization of both the extrinsic and the intrinsic elements of form.

strength of diplomatics – its being "rooted in jurisprudence, administrative history and theory and a body of historical and contemporary knowledge about the nature of record-keeping practices in bureaucratic organizations" (MacNeil, 2004, p. 224) – may also be seen as a limitation, if one considers the diversity of environments in which documents and other information artifacts are created. If we are to understand the interactions taking place in today's workplaces as well as in any other contexts of document production, including unstructured and 'juridically irrelevant' ones, we need to build new knowledge from the direct observation of "living information systems".

Genre theory, with its notion of *genre evolution*, its broader sphere of interest, and its focus on the "negotiations among social actors that results in shared typifications which gradually acquire the moral and ontological status of taken-for-granted facts" (Yates and Orlikowski, 1992, p. 305), can usefully complement the traditional diplomatic model, in order to make the latter more dynamic and better responsive to the transformations occurring in society. Diplomatics did emerge through an "inductive process of observation and comparison" (MacNeil, 2004, p. 225); however, early diplomatists had access to a relatively limited number of surviving documents and could not observe the process of document creation as it was carried out. Genre theory provides a set of tools to elicit information from participants in the creation process (Spinuzzi, 2003). Understanding the motivations, "exigencies" and unexpressed needs behind the creation and modification of documentary products is a pre-requisite to engage in the process of systematization and abstraction necessary to develop models of contemporary documents.

Diplomatics provides conceptual and terminological rigor to discussions of digital objects, as well as a fundamental lesson: in order to understand and evaluate the functions and intrinsic meaning of the documentary residue of activities, one should directly examine the documents embodying those activities and identify the purpose for which they were created.

## 4 The Impact of Genre on Information Retrieval

In the last several decades, the information retrieval research community has developed a tradition of encouraging statistical analysis of lexical content as the dominant driving force behind many information search, navigation and seeking activities, e.g. bag-of-words, *term frequency-inverse document frequency* (TF-IDF), language models.[3] Some of these methods have also been presented as viable methods for non-textual media.[4]

The focus on lexical content analysis in the research community is remarkable, given that relevance ranking Google, of one of the most popular search engines[5], has relied for many years on the link structure between pages (as opposed to relying solely on the presence of specific terms or phrases within the pages). The basic assumption is that those pages with many in-coming links (i.e. pages to which other pages refer often) are likely to be more important than others. The effectiveness of Google's method in satisfying

[3] See Manning et al. (2009).

[4] See Magalhaes and Rueger (2007).

[5] See at `http://searchenginewatch.com/reports`.

the expectations of a substantial proportion of searchers a large proportion of the time is indicative of a crucial aspect of information: the value of information is often measured by its context and frequency of use.

There has been a recent growth of attention in the information retrieval community toward context of information use and human interaction behavior.[6] This attention is driven by a variety of factors, including the desire to disambiguate queries as reflections of information needs[7], differences arising from personal requirements[8] and varying levels of complexity of distinct tasks[9] – as evidenced by new tracks within the Special Interest Group on Information Retrieval (SIGIR) of the Association for Computing Machinery (ACM).[10]

Document representation models can be based on two distinct perspectives (Mehler et al., 2010a): one focused on their *meaning* (or content) and the other focused on their genre-related *function*. In order to illustrate these two dimensions, consider the example of personal academic homepages (Rehm, 2002). Instances of this webgenre serve a relatively stable set of functions in that they inform, e.g., about the CV, the publications, projects and teaching duties of the author or owner of the site). Their content, however, can vary significantly based on a variety of factors, including the disciplinary affiliation of the creator. In many other cases (e.g. weather reports), both the type of content and functions of the content are relatively consistent and stable over time. Following this line of reasoning, any document collection can be made accessible by its content or by its function, where content is typically modeled in the framework of the bag-of-words model (Salton, 1989), while function can be modeled, for example, in terms of a bag-of-structures model (Mehler and Waltinger, 2010) or by means of bags of features that map lexical manifestations of functional units (Santini, 2010; Sharoff, 2010).

Genre modeling can provide valuable guidance about the contextual parameters that determine, select, modify or otherwise constrain the functions associated with a document to be preserved. Let us return to our earlier example. Changes in the capabilities and affordances of supporting information technology have significantly shaped the functions being performed by academic homepages. Digital objects can be persistently identified through usually a variety of naming and resolution mechanisms, and systems – CiteULike (Hammond et al., 2005) and BibSonomy (Hotho et al., 2006) – can dynamically cluster documents by exploiting various relationships implied by common data values in associated tags or other metadata. Because social tagging is time-dependent, it can reflect the changing relationships between documents and associated scientific communities, based on the semantic assertions implied by the tags. The personal academic homepage can thus take on a new function, viz. locating one's work within a larger academic discourse. This is an emergent function whose final gestalt is not yet clear.

Users of preserved digital objects can benefit from not only information about the

[6] See at `http://www.iiix2010.org/`.

[7] See Stojanovic (2005).

[8] See at `http://personal.cis.strath.ac.uk/~ir/accessiblesearch/final_proceedings.pdf`.

[9] See Wu et al. (2008).

[10] See at `http://portal.acm.org/citation.cfm?id=1835449&coll=ACM&dl=ACM&CFID=101627353&CFTOKEN=54139786`.

| | closed | | open | |
|---|---|---|---|---|
| | time point-related | time period-related | time period-related | course of time-related |
| **topic** | | | | |
| micro-level topic | | | | |
| meso-level topic | *content classification scheme* | | *model of emergent topics* | |
| macro-level topic | | | | |
| **function** | | | | |
| micro-level genre | | | | |
| meso-level genre | *genre palette* | | *model of emergent genres* | |
| macro-level genre | | | | |

Table 1: Topic and function of documents in relation to closed and open document models (Mehler and Waltinger, 2009) on the level of documents (meso-level), their constituents (micro-level) and document networks (macro-level).

contextual dynamics of the objects themselves (e.g. how and when they were created, how they have been transformed, how they have been used) but also from information about the contextual dynamics of descriptive information associated with the objects (e.g. how and when metadata elements were created, how they have been transformed, how they have been used). Descriptors are contextualized, for example, by the individuals who tagged them, by their "validity period" and by the network of documents into which the documents are linked. This suggests that models for descriptive information in long-term preservation environments should include time variables in order to capture the life cycle of digital genres, their validity periods and rates of change.

Classically, the role of genre in information retrieval has been identified with providing a further retrieval dimension (Rehm, 2002) whose structural manifestation may be used to arrive at more effective document representations (Lindemann and Littig, 2010). Digital preservation activities will additionally benefit from an *open document model* that reflects the dynamics of document representations (Mehler and Waltinger, 2009):

1. A document model is said to be *closed* if its composition is fixed or almost fixed. A closed model is given by a classification scheme – e.g. Dewey Decimal Classification (DDC) or Medical Subject Headings (MeSH) – that provides a system of thematic or functional categories of a certain area of application. Typically, a closed model is generated by a small number of experts (or even by a single expert as in the case of many genre palettes). From the point of view of a closed model, the representation of a document is fixed – irrespective of when the model is applied to the document. Typically, closed (topic or genre) models are implemented in the framework of supervised learning (Sebastiani, 2002). After phases of training and testing, the model is applied until it is replaced by a follow-up model. Here, the classification scheme specifies the target categories to be learned.

2. A document model is said to be *open* if its composition is in a state of permanent flux according to the temporal dynamics of its application area. Typically, an open model is generated by a large community of possibly many non-experts who cooperate in

a self-organized manner without central supervisors. An example of an open topic model is the social ontology (i.e., category system) of Wikipedia (Mehler, 2010). From the point of view of an open model, the representation of a document is fluid: it depends on the time of its application to the document. Open topic models can be implemented in the framework of unsupervised learning that does nor require a formal training phase.

By additionally distinguishing between the micro-level of document constituents and the macro-level of document networks we finally get a matrix of document representation levels as given in Table 1.

*So how does a document model look like that captures these additional dimensions?* Starting from an open topic model $\mathcal{T}$ as given by Wikipedia's category system, such a document model can be outlined in terms of bags-of-features as follows: a document $x$ to be preserved at time $t = 1$ is represented by a time series

$$(T_1, \ldots, T_n)$$

of bags of features $T_i$, $i \in \{1, \ldots, n\}$, such that for any time point $1 \leq i \leq n$, $T_i$ is a mapping (i.e, a set of thematic descriptors) that locate $x$ in the social ontology $\mathcal{T}_i$ at time $i$. Suppose that the same is done by the time series

$$(G_1, \ldots, G_n)$$

for some open genre model $\mathcal{G}$ (see Table 1). Now, if document $x$ is accessed at time $n$, a user can "reactivate" the contextual embedding of $x$ at any preceding time back to the initial time of preservation. That is, she can select not only thematic or generic descriptors, but also a point or period of time that contextualizes her search query. Moreover, the user may also ask the system to "translate" her search query so that it fits more appropriately to a past state of the semantic universe from which $x$ was selected. In this way, it should be possible to cope with semantic change due to descriptors that have been used to tag $x$ at $t = 1$, and have changed their meaning in the meantime. Evidently, such a representation model asks for a periodic renewal of the representation of $x$ by means of thematic or generic tags.

From our point of view, a retrieval model of this kind is needed to cope with the dynamics of document representation in long-term digital preservation situations. Documents can then be re-encountered within the context of a previously activated set of semantic relationships – even after decades of language change, topic change and genre change. This is an instance of the broader goal of carrying contextual information over time through persistent state information (Lee, 2010).

## 5 Next Steps

In this paper, we emphasized the central role that contexts of use and underlying human activities play in establishing effective value-added digital preservation, archival practices,

and information retrieval. It seems crucial that future research in automation of digital preservation should include investigation into automation processes that can reflect aspects of the context of objects at various points throughout their life-cycle.

As a consequence of our observations, we offer the following potential questions for future research related to the automation of digital preservation:

- What are the limits of functional information context that can be automatically extracted from selected objects, collections, social networking environments (e.g. wiki and blogs)?
- What kind of contextual information should be retained to optimize digital preservation within environments in which continued retention of content cannot be guaranteed?
- What can archival science teach the information retrieval research community with respect to finding evidence of valuable information?
- What new types of human transactions arise from information seeking processes such as information retrieval? Do these affect archival practices in documenting evidence of human activities?
- What contextual factors should be preserved in order to improve search, navigation, seeking, browsing, and discovery in the future?
- Can we feed information about information use back into the preservation framework to create an automatic adaptive digital preservation system to support information usage?

As a first point of exploration, we have drawn attention to previous research areas that recognize a continuously evolving range of communicative genres as social manifestations of human activities in relation to managing, using and creating information. This suggests further investigations of identifying the potential of genre as a gateway to valuable context.

Previous work has proposed investigation into the automation of genre classification to support metadata extraction (Kim and Ross, 2006) and to facilitate such archival practices as appraisal based on genre identification and metadata extraction (Oliver, Kim and Ross, 2008). This position paper shows that by looking at the issue from a broad range of digital preservation disciplines a wider array of research domains emerge which will enhance preservation processes. Appropriate methodologies will be similarly varied, ranging from ethnographic investigation to document analysis. Specific research questions include:

- *What can archival concepts in general and diplomatics in specific contribute to the study of genre?*
- *What can the notion of genre contribute to archival studies (including diplomatics)?*
- *Can genre contribute to assessments authenticity of digital objects?*

- *What can genre tell us about the usage of digital objects?*
- *What can genre tell us about the intentions of document creators?*
- *What can genre tell us about the agency of the creator or the user (whether human or machine)?*
- *How much contextual information is needed to guarantee an understanding of digital object in the future?*
- *How far can we go in modeling genre as a social-semiotic concept in terms of text-based machine learning?*
- *To what degree are genre and topic (i.e. function and content) orthogonal categories? How do they interact?*
- *How can we integrate genre-related constituents (e.g. diplomatic (sub-)procedures) into a genre-sensitive retrieval model and genre-sensitive search queries?*
- *How does one reflect genre over time, given both (1) the inability to reflect all aspects of context and (2) changes to the underlying technologies used to encode and represent digital objects?*
- *What kinds of experimental environments or apparatus should we put in place to enable the pursuit of this area of research (e.g. there is a need for corpora to support genre research.*[11]

The authors of this paper intend to apply the concepts and methods described above in a number of different contexts. One promising area of application is the preservation of social software.

### Acknowledgement

We gratefully acknowledge the inspiring and constructive atmosphere at the Dagstuhl seminar *Automation in Digital Preservation* as well as all fruitful discussions with our colleagues.

## References

Barbiche, B. (1996). Diplomatics of modern official documents (sixteenth-eighteenth centuries): Evaluation and perspectives. *The American Archivist*, 59(4):422–436.

Barwise, J. and Perry, J. (1983). *Situations and Attitudes*. MIT Press, Cambridge.

Bateson, G. (1972). *Steps to an ecology of mind: collected essays in anthropology, psychiatry, evolution, and epistemology*. Chandler Publishing, San Francisco.

[11] See for instance Berninger et al. (2008) or Rehm et al. (2008).

Bawarshi, A. (2001). The ecology of genre. In Weisser, C. R. and Dobrin, S. I., editors, *Ecocomposition: Theoretical and Pedagogical Approaches*, pages 69–80. SUNY Press, Albany.

Berninger, V. F., Kim, Y., and Ross, S. (2008). Building a document genre corpus: a profile of the KRYS I corpus. In *Proceedings of the BCS-IRSG Workshop on Corpus Profiling, London, 18th October 2008*.

Boyle, L. E. (1976). Diplomatics. In Powell, J. M., editor, *Medieval Studies: An Introduction*, pages 69–101. Syracuse University Press, Syracuse, NY.

Cerviño Beresi, U., Kim, Y., Baillie, M., Ruthven, I., and Song, D. (2010). Relevance in technicolor. In Lalmas, M., Jose, J. M., Rauber, A., Sebastiani, F., and Frommholz, I., editors, *Proceedings of the 14th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2010), Glasgow, UK, September 6-10*, volume 6273 of *LNCS*, pages 196–207. Springer, Berlin.

Duranti, L. (1998). *Diplomatics. New Uses for an Old Science*. SAA and ACA in association with the Scarecrow Press Inc., Lanham and London.

Guyotjeannin, O. (1996). The expansion of diplomatics as a discipline. *American Archivist*, 59(4):414–21.

Hammond, T., Hannay, T., Lund, B., and Scott, J. (2005). Social bookmarking tools (i): A general review. *D-Lib Magazine*, 11(4).

Hjelmslev, L. (1969). *Prolegomena to a Theory of Language*. University of Wisconsin Press, Madison.

Hotho, A., Jäschke, R., Schmitz, C., and Stumme, G. (2006). BibSonomy: A social bookmark and publication sharing system. In *Proc. of the Workshop on Tool Interoperability at the International Conference on Conceptual Structures 2006*, pages 87–102.

Kim, Y. and Ross, S. (2006). Genre classification in automated ingest and appraisal metadata. In Gonzalo, J., Thanos, C., Verdejo, M. F., and Carrasco, R. C., editors, *Research and Advanced Technology for Digital Libraries, 10th European Conference, ECDL 2006, Alicante, Spain, September 17-22*, number 4172 in LNCS, pages 63–74, Berlin/New York. Springer.

Kim, Y. and Ross, S. (2007a). Detecting family resemblance: Automated genre classification. *CODATA Data Science Journal*, 6:172–183.

Kim, Y. and Ross, S. (2007b). Searching for ground truth: a stepping stone in automated genre classification. In Costantino Thanos, Francesca Borri, L. C., editor, *Proceedings of First International DELOS Conference on Digital Libraries: Research and Development, Pisa, Italy, February 13-14, 2007*, number 4877 in LNCS, pages 248–261, London, UK. Springer.

Kim, Y. and Ross, S. (2007c). "The naming of cats": Automated genre classification. *International Journal of Digital Curation*, 2(1):49–62.

Lee, C. (2010). A framework for contextual information in digital collections. *Journal of Documentation*, 67(1).

Lindemann, C. and Littig, L. (2010). Classification of web sites at super-genre level. In Mehler et al. (2010b).

MacNeil, H. (2004). Contemporary archival diplomatics as a method of inquiry: Lessons learned from two research projects. *Archival Science*, 4:199–232.

Magalhaes, J. and Rueger, S. (2007). High-dimensional visual vocabularies for image retrieval. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 815–816, New York, NY, USA. ACM. `http://doi.acm.org/10.1145/1277741.1277923`.

Manning, C. D., Raghavan, P., and Schütze, H. (2009). *Introduction to Information Retrieval*. Cambridge University Press, Cambridge. `http://nlp.stanford.edu/IR-book/html/htmledition/irbook.html`.

Mehler, A. (2010). A quantitative graph model of social ontologies by example of Wikipedia. In Dehmer, M., Emmert-Streib, F., and Mehler, A., editors, *Towards an Information Theory of Complex Networks: Statistical Methods and Applications*. Birkhäuser, Boston/Basel.

Mehler, A., Kühnberger, K.-U., Lobin, H., Lüngen, H., Storrer, A., and Witt, A., editors (2010a). *Modeling, Learning and Processing of Text Technological Data Structures*. Studies in Computational Intelligence. Springer, Berlin/New York. In preparation.

Mehler, A., Sharoff, S., and Santini, M., editors (2010b). *Genres on the Web: Computational Models and Empirical Studies*. Springer, Dordrecht.

Mehler, A. and Waltinger, U. (2009). Enhancing document modeling by means of open topic models: Crossing the frontier of classification schemes in digital libraries by example of the DDC. *Library Hi Tech*, 27(4).

Mehler, A. and Waltinger, U. (2010). Integrating content and structure learning: A model of hypertext zoning and sounding. In Mehler, A., Kühnberger, K.-U., Lobin, H., Lüngen, H., Storrer, A., and Witt, A., editors, *Modeling, Learning and Processing of Text Technological Data Structures*, Studies in Computational Intelligence. Springer, Berlin/New York.

Miller, C. R. (1984). Genre as social action. *Quarterly Journal of Speech*, 70:151–167.

Morris, C. W. (1938). *Foundations of the Theory of Signs (International encyclopedia of unified science)*. Chicago University Press, Chicago.

Østerlund, C. (2007). Genre combinations: A window into dynamic communication practices. *Journal of Management Information Systems*, 23(4):81–108.

Peirce, C. S. (1934). *Pragmatism and Pragmaticism*, volume V of Collected Papers of Charles Sanders Peirce. Harvard University Press, Cambridge.

Pratesi, A. (1962). *Elementi di diplomatica generale*. Adriatica Editrice, Bari.

Rehm, G. (2002). Towards automatic web genre identification – a corpus-based approach in the domain of academia by example of the academic's personal homepage. In *Proc. of the Hawaii Internat. Conf. on System Sciences*.

Rehm, G., Santini, M., Mehler, A., Braslavski, P., Gleim, R., Stubbe, A., Symonenko, S., Tavosanis, M., and Vidulin, V. (2008). Towards a reference corpus of web genres for the evaluation of genre identification systems. In *Proceedings of the 6th Language Resources and Evaluation Conference (LREC 2008), Marrakech (Morocco)*.

Salton, G. (1989). *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison Wesley, Reading, Massachusetts.

Santini, M. (2010). Cross-testing a genre classification model for the web. In Mehler et al. (2010b).

Santini, M., Mehler, A., and Sharoff, S. (2010). Riding the rough waves of genre on the web: Concepts and research questions. In Mehler et al. (2010b), pages 3–32.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47.

Sharoff, S. (2010). In the garden and in the jungle. Comparing genres in the BNC and Internet. In Mehler et al. (2010b).

Spinuzzi, C. (2003). *Tracing Genres through Organizations. A Sociocultural Approach to Information Design*. The MIT Press, Cambridge.

Stojanovic, N. (2005). Information-need driven query refinement. *Web Intelli. and Agent Sys.*, 3(3):155–169. `http://portal.acm.org/citation.cfm?id=1239800`.

Wittgenstein, L. (1953). *Philosophical Investigations*. Blackwell Publishing, Oxford.

Wu, I.-C., Liu, D.-R., and Chang, P.-C. (2008). Toward incorporating a task-stage identification technique into the long-term document support process. *Inf. Process. Manage.*, 44(5):1649–1672. `http://dx.doi.org/10.1016/j.ipm.2007.11.005`.

Yates, J. and Orlikowski, W. (2002). Genre systems: Structuring interaction through communicative norms. *Journal of Business Communication*, 39(1):13–35.

Yates, J. and Orlikowski, W. J. (1992). Genres of organizational communication: A structurational approach to studying communications and media. *Academy of Management Review*, 17(2):299–326.

Yoshioka, T., Herman, G., Yates, J., and Orlikowski, W. (2001). Genre taxonomy: A knowledge repository of communicative actions. *ACM Transactions of Information Systems*, 19(4):431–456.