Toward a 5M Model of Digital Libraries

Sándor Darányi
Swedish School of Library and Information Science
University of Borås
sandor.daranyi@hb.se

Peter Wittek
Department of Computer Science
National University of Singapore
wittek@nus.edu.sg

Milena Dobreva
Centre for Digital Library Research
University of Strathclyde
milena.dobreva@strath.ac.uk

ABSTRACT
Whereas the DELOS DRM and the 5S model of digital libraries (DL) addresses the formal side of DL, we argue that a parallel 5M model is emerging as best practice worldwide, integrating multicultural, multilingual, multimodal digital objects with multivariate statistics-based document indexing, categorization and retrieval methods. The fifth M stands for the modeling the information searching behavior of users, and of collection development. We show how an extension of the 5S model to Hilbert space (a) points toward the integration of several Ms; (b) makes the tracking of evolving semantic content feasible, and (c) leads to a field interpretation of word and sentence semantics underlying language change. First experimental results from the Strathprints e-repository verify the mathematical foundations of the 5M model.

Categories and Subject Descriptors
G.1.2 [**Approximation**]: Wavelets and fractals; H.3.7 [**Digital Libraries**]: Systems issues, User issues.

General Terms
Measurement, Experimentation, Human Factors, Theory.

Keywords
5S model; 5M model; Hilbert space; wavelet analysis; text categorization.

1. INTRODUCTION
Presenting here work in progress from the SHAMAN EU FP7 Integrated Project, we expect radical changes in DLs in the following three areas: (1) Expanding the 5S model of DLs [19] to include potentially infinite dimensional Hilbert spaces, leading to the use of new, more sophisticated methods beyond Support Vector Machines [21], for the indexing, categorization and retrieval of digital objects. This extension renders a physical interpretation of vectors and functions possible to keep track of the evolving semantics and usage context of the digital objects, including metadata and workflow evolution; (2) Utilization of the higher information representation capacity of mathematical objects in the above spaces. New ways of usage will lead to the integration of word and sentence semantics in document categorization and retrieval models. Physics as a metaphor will play the role of an interface between language and mathematics, rendering words as specific mass and/or energy values, and sentences, documents queries and states of databases as their superpositions; (3) The above create ways for language representation for optical and quantum computing, leading to document processing applications by environmentally friendly supercomputing.

These changes will motivate the DL community to add a highly integrated 5M layer to the 5S formal model of DL, to be introduced below, where 5M stands for DL collections with Multicultural, Multilingual, Multimodal content; indexed, categorized and retrieved by Multivariate algorithms; and harnessing their evolving semantics by the Modeling of collections and users.

This paper is structured as follows: in Section 2, we discuss state of the art DLs with an eye on the DELOS Digital Library Reference Model (DLRM); in Sections 3 and 4, components of the 5M model and their anticipated connections to 5S and DLRM are discussed. In Section 5, we extend the 5S model by Lebesgue spaces. In Section 6, a field interpretation of semantics is adopted for vector and function spaces, with a brief discussion of wavelets and the proposed new semantic kernel in Section 7. Section 8 presents a first case study with its discussion in Section 9, and the conclusions and future work in Section 10.

2. STATE OF THE ART
Modern DLs have to address multiple diverging requirements and expectations accommodating also the rapidly changing technological novelties. Currently it seems that DLs are built basically to fill in a specific gap of provision[1] and/or to find out how a specific technological solution could improve the functionality of a DL. As Michael Khoo *et al*. noted [22]:

> *"In the case of digital library researchers, the focus of research is often on technical issues (e.g., information retrieval methods, software architecture, etc.) rather than on user-centered issues."*

This reinforces the need to look at DLs from a more holistic point of view, of special importance when they are built and evaluated for international development. Such more holistic perspectives were already addressed when models such as DELOS DLRM [6] and the 5S model had been developed.

DELOS DLRM analyses six basic domains in the digital library universe: Content, User, Functionality, Architecture, Policy and

Conference'10, Month 1–2, 2010, City, State, Country.

[1] For example most of the current project related to Europeana (www.europeana.eu) address specific types of content, e.g. Musical Instrument Museums Online (MIMO) (http://www.mimo-project.eu/) aims to create "a single access point to digital content and information on the collections of musical instruments held in European museums."

Quality. It introduces a number of concepts from these domains which are also seen as related to one another. This actually allows to study in more depth one domain or to see what the inner links between domains are: for example how quality parameters of DL content influence user satisfaction or what functionality features are especially attractive or disliked by users.

According to DLRM every domain in the DL universe is decomposed into a finer level of granularity and its central concept is defined. In the Content domain this is the information object [6]:

> *The **Content** concept encompasses the data and information that the Digital Library handles and makes available to its users. It is composed of a set of information objects organised in collections. Content is an umbrella concept used to aggregate all forms of information objects that a Digital Library collects, manages and delivers. (p. 19])*

This structural view is an excellent illustration of the observation of Simon Tanner [30]:

> *"Digital libraries are trying to move from managing **containers** and **content** to managing **context** and it is proving to be a larger and historically more difficult challenge to overcome." (p. 38-39)*

Containers here are meant to be the physical carriers of information and could be compared with the Resources in the DLRM. Content as used by Tanner is not identical to the generic domain of Content but is closer to the concept of Information object in DELOS DLRM. With managing context as the great challenge in the domain and consequently calling for models able to address its complexity, our conceptual framework is the generic model of context by [15, 5].

The fact that DLs are used in an international environment means that they should address properly **multicultural** and **multilingual** issues. For example, a recent international Europeana user and functionality study showed that participants across four countries expected multilingual support [12]. But multilinguality was not a stand-alone requirement since participants also wanted to be able to see the digital objects in context which would help to understand better their cultural setting. With the range of document **modalities** in a typical DL, one can argue that without the introduction of a uniform methodology to process them, no high-level integration between form and content can be accomplished; and that such collections embody mankind's best chance this far to simulate process outcomes whose components are digital objects. To address these concerns however requires a first draft of a new integrating model plus in accord with it a new way for information representation.

While these three concepts – the multiculturality, multilinguality and multimodality – of a digital library characterize the current international and technological dimensions of their use, there is also the challenge of applying such models to DLs which will be able to address the federated nature of resources and their very large size in terms not only of volume of digital objects but their numbers as well. From this perspective we believe that the concept of Content as defined by DLRM can be refined further to address not only the single objects but also their characteristics and the appropriate issue of size.

Here one could argue that this had already been done in the 5S model. This model introduces the basic notions of Streams, Structures, Spaces, Scenarios and Societies. The model is very useful to formalize the representation of objects in digital libraries, but again the context within the digital libraries remains unaddressed. This motivated us to suggest a new model, called 5M, which looks in more detail into the specific characteristics of current digital libraries crystallizing a point of view which helps to position the notion of context.

This model should not be seen as "yet another model" of digital libraries but as something which builds a bridge between DELOS DLRM and 5S addressing deeper the characteristics of the modern DLs and taking into account not only containers and content, but also context.

3. THE *5M* MODEL: A NEW PERSPECTIVE ON CONTEXT IN DIGITAL LIBRARIES

In DLs with Multicultural, Multilingual, Multimodal documents, plus their content processed by Multivariate statistical algorithms, adding the Modelling of user behaviour and content evolution completes what we call the 5M model. Different blends of the first four Ms already exist in DLs worldwide, with user modeling taking off as well [11], plus e.g. simulations based on DL data [24]. However, their conscious integration into a 5M model is yet to take place. Matching 5S with 5M will, in our eyes, lead to a next level of integration between form and content.

3.1 Multicultural dimension

Relating or pertaining to several different cultures, multiculturalism is the doctrine that several different cultures (rather than one national culture) can coexist peacefully and equitably in a single country, accepting or promoting multiple ethnic cultures, for practical reasons and/or for the sake of diversity and applied to the demographic make-up of a specific place, usually at the organizational level, e.g. schools, businesses, neighborhoods, cities or nations, extending equitable status to distinct groups without promoting any specific ethnic, religious, and/or cultural community values as central. In a DL context, this principle reflects the preservation of different cultures or cultural identities within a unified society, as a state or nation.

3.2 Multilingual dimension

Multilingualism being the use of two or more languages either by an individual speaker or by a community of speakers, by multilingual content in DLs we define digital objects concurrently indexed by and/or accessible in more than one language. A typical implementation of this principle is cross-language text categorization and/or information retrieval. Multilinguality of DLs was identified as an issue over a decade ago [27] but the solutions to support multilingual search are still not universal and widely used. The CACAO project[2], for example, worked on "innovative approach for accessing, understanding and navigating multilingual textual content in digital libraries and library catalogues"; however restricted to subject domains such as mathematics, geography and history.

[2] http://www.cacaoproject.eu/

3.3 Multimodal dimension

This dimension pertains to documents (digital objects) as they are stored in DLs, i.e. their form as opposed to their content. Typical document modes are text, image, video and audio.

3.4 Multivariate dimension

Multivariate or multidimensional refers to an entity described by more than one of its features, formalized as variables, where these can have two or more discrete or continuous values. For example text features can be the content words in a document; or, in an image retrieval context for DLs, image content may refer to colors, shapes, textures, or any other information that can be extracted from the image itself. Without the ability to examine image content, searches must rely on metadata such as captions or keywords, which in turn can be treated as multivariate descriptions of the digital object.

3.5 Modeling dimension

A model is a hypothetical description of a complex entity or process. More specifically, a *mathematical model* is 'a representation of the essential aspects of an existing system (or a system to be constructed) which presents knowledge of that system in usable form' (17), used in natural science, engineering and the social sciences (economics, psychology, sociology, and political science).

4. THE *5M* MODEL AND ITS CONNECTION TO DELOS DLRM and *5S*

Components of our 5M model clearly connect to one or more elements of DLRM or the 5S model (Streams, Structures, Spaces, Scenarios, Societies). A formal expression of such connections will have to be addressed by future research.

As for DLRM, multicultural and multilingual aspects of digital objects fall under Content, multimodal content expressed and processed by multivariate methods including its modeling under Functionality, and the aforementioned as implications of diverse information needs under Users.

In 5S, Streams cover multimodality, Structures imply document classifications, Spaces include multivariate methods and their use to process cross-language documents in vector space, Scenarios as sequences of events with specific parameters allow for modeling, and Societies represent users as well.

However, in the 5S model, natural language (NL) as a formal component is not considered [19]. On the other hand, the crux of developing well-functioning domain-specific sublanguages of concept spaces for DL requires a fit between selected properties of NL and the capacities of mathematical objects to model them. Therefore we propose to augment the set of spaces within the 5S model by function (Hilbert) spaces [36], enabling a physical interpretation of vectors and functions. Moving away from representing terms, documents and queries exclusively by linear algebra removes some of the barriers of modeling NL and results in a more encompassing view, still in accord with the 5S model. Such an interpretation is permitted by the definition of SVM [31] for text categorization (TC). We demonstrate this by a new semantic kernel based on signal processing using wavelets. This also paves the way for a joint representation of images and text, addressing multimodality and semantics from a different angle.

5. EXTENDING THE *5S* MODEL BY L^2 (LEBESGUE) SPACES

Following [19], we regard spaces as a formal component of digital libraries. While different kinds of spaces, such as probability spaces and high-dimensional real vector spaces, have been widely used in a DL context, there are other spaces that can be useful. In this paper we argue in favor of the L^2 linear measure space.

Let $1 \leq p \leq \infty$ and (S, Σ, μ)be a measure space. Consider the set of all measurable functions from S to $\mathbb{C}$ whose absolute value raised to the p-th power has a finite Lebesgue integral. The set of such functions forms a vector space, with the usual natural vector operations of addition and multiplication by scalars. This can be made into a normed vector space $L^p(S, \mu)$, and for $p = 2$, this will be a Hilbert space. This means that the L^2 space is equipped with an inner product just like the dot product of the Euclidean space. The advantage of the L^2 space is that it allows exploiting the dependency between the index terms in a computationally tractable manner. Further, changing the type of space enables a richer document and database structure addressing evolving semantics, where the processing of located content, and vectors as handles to it, is methodologically congruent with that of dislocated (evolving) content.

We expect these two shifts to affect Streams and Services directly and Scenarios indirectly. As [2] defines it, "[d]ynamic streams facilitate communication in digital libraries". In this regard, filtering of documents and document ranking may be considered Streams, and thus will be influenced by the underlying mathematical representation of digital objects. Scenarios describe the functionalities of digital libraries by combining the other components, hence a change in Spaces and Streams will have an indirect impact on Scenarios and Services as well.

6. VECTOR AND FUNCTION SPACE REINTERPRETED: A FIELD THEORY OF WORD SEMANTICS

Extending the 5S model to include L^2 space is a convenient step to develop a better understanding of word semantics, and thereby an improved modeling of the semantics of full texts as well. The necessary steps to build a comprehensive frame of thought requires a sometimes radical reinterpretation of some conceptual tools and the ways we have been accustomed to use them, because of discrepancies between texts in NL and its vector representation. On the other hand, the gap between language and mathematics can be bridged by physics. Our argumentation is as follows:

1. Previous research suggests that the efficiency rate of wavelet-based text categorization (TC) models is often better than that of vector models [33]. This indicates that semantic content is at least as much conform with its function representation as with its vector representation. This situation is reminiscent of the particle- wave duality in physics [4]. The explanation for the competitiveness of wavelets can be only the fact that an exclusively mass point like interpretation of semantic content is incorrect;
2. The dynamism of evolving semantics requires two types of vectors, not one, that is, vectors from physics, not linear algebra. Vectors as mathematical objects do not per se correspond to content: as language has no center,

any origin of a coordinate system crucial for vector space models is simply not there, therefore vectors cannot be genuine representations of e.g. terms; likewise, vector length has no intuitive explanation as being important for word semantics. However, things fall in their place as soon as we regard vectors as pointers at term etc. locations, i.e. see them in a different role: they do not embody content but point at it. This role is absolutely necessary, because one cannot add up etc. points, only by somehow identifying them, where a vector is the means of identification. In other words, we can keep vectors (n-grams) as mathematical illustrations of content behavior as long as we shift of our focus from linear algebra to physics, where both position and direction vectors are routinely used to model e.g. field behavior;

3. Why do wavelets perform in some cases better than vectors? The answer is, if semantic content is not exclusively mass point like then it cannot be exactly localized [16, 33] but "smeared out", very much like electrons have no exact positions, only a probability distribution indicating how likely it is to find a certain charge at a certain location in a certain moment. This regional nature of semantic content requires regional pointers, such as wavelets, and not exact ones like vectors, moving away from the picture of a term charged with meaning as a mass point with exact location;
4. The regional nature of semantic content confirms what has been known as semantic fields or lexical fields in linguistics [25]. Adding dynamics to the picture outlines characteristics of a physical field, one that demands a very different model from our current thinking. Therefore we regard semantic content as a kind of charge or gravity of some geometric vehicle such as a location [9], to be displaced over time. This view follows the flow model of [14] in the general probabilistic framework of language change [2];
5. Applying wavelets to model word meaning only stresses the benefits of the physical interpretation of semantic spaces. Wavelets being mathematical abstractions localized both in time and frequency, documents and queries, as much as they are sums of term vectors in the VSM, become wavelet sums. Such sums are called interference in optics and superposition in mechanics. What matters though is that the sum of locations remains still a location.

The above observations persuade us to treat word semantics as a field, a field being a spatiotemporal distribution of some quality (e.g. charge, mass) quantified by its values and represented e.g. as vector space with position and direction vectors. In DL, typically two sources of word meaning are used for indexing: distributional similarity (contextuality) and reference. Contextuality drives the VSM, while reference plays a role in lexical resource-based measures when these are coupled e.g. with the VSM, where the lexical resource is the external component charging a term location, its referent. Such a coupling satisfies the above field definition. Furthermore, in L^2 space where wavelet descriptors are deduced from document signals based on the semantic ordering of terms, a procedure that charges information on term occurrence with ontological content, the integration of both sources of word semantics in one representation results in a field in a quasi-physical sense. Rules underlying the behavior of this evolving field constitute what we term the field theory of word semantics.

7. WAVELETS AND THE CSBF KERNEL

With the mathematical details reported in [34], in this frame of thought we used compactly supported basis functions (CSBF) as Support Vector Kernels. The proposed CSBF kernel heavily utilizes term interdependence, with related features are assigned to subsequent vectors of the canonical base of the L^2 space. To this end, conceptually related features are grouped together based on their one-dimensional semantic ordering (1, 33). This procedure utilizes word sense relations from WordNet (18), "charging" term locations with word semantics and creating thereby a field. Further, we assume that if a set of features is reordered according to some relatedness measure, then the vector representation of an object can be regarded a series of equally spaced observations of a continuous signal, as if it were a time series but constructed by statistical relatedness, and reconstructed by the Whittaker-Shannon-formula (32).

Assuming that the related features f_{i-1}, f_i and f_{i+1}follow each other, consider the following example. The first object has the feature f_i, and so does the second object. In Figure 1 it can be seen that feature f_i is counted the same way as it would be in a vector space model, the related features f_{i-1} and $f_{i+1}$are counted to a smaller extent, while other related features are considered even less. If, on the other hand, the two objects do not share the exact feature and only related features occur, for instance f_{i-1} and f_{i+1}respectively, then the feature f_i, placed between f_{i-1} and f_{i+1} in the same order, will be considered to some extent for the calculation of similarity (Figure 2).

8. PUTTING *5M* TO WORK: A CASE STUDY

In order to test how our model can be applied in real life, we conducted an experiment with the Strathprints digital repository. This is an institutional e-print repository for making research papers and other scholarly publications widely available on the Internet at the University of Strathclyde, UK, its hosting and technical support provided by the Department of Computer and Information Sciences (CIS) [10]. E-prints and usage statistics software have been installed, configured and managed by the Centre for Digital Library Research (CDLR) at the same

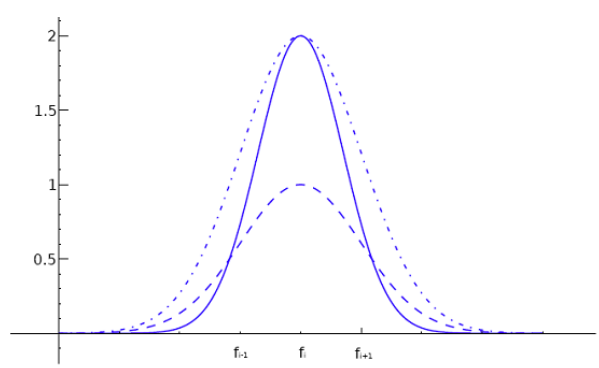
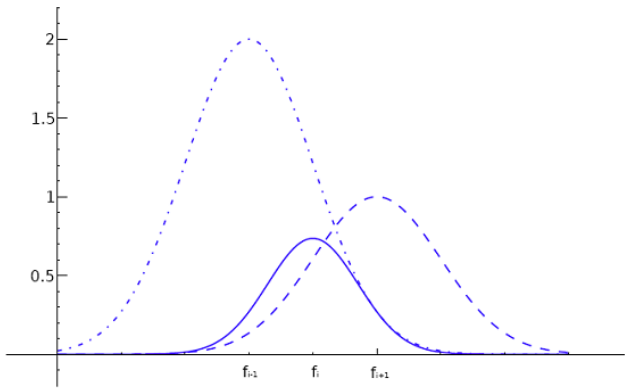


Figure 1. Two objects with a matching feature f_i. Dotted line: Object-1. Dashed line: Object-2. Solid line: their product.



. Figure 2. Two objects with no matching features but with related features f_{i-1} and f_{i+1}. Dotted line: Object-1. Dashed line: Object-2. Solid line: their product

university. Its digital objects are indexed by the LCSH classification scheme. From a machine learning (ML) point of view, this is a multilabel scenario where an instance of the collection may belong to several categories. We downloaded and processed 5946 abstracts with LCSH metadata. Keywords were obtained by a WordNet-based stemmer using the controlled vocabulary of the lexical database resulting in 21718 keywords in the full-text documents and 11586in the abstracts. Keywords were ranked according to the Jiang-Conrath distance [20] based with the algorithm described in [33].

With 20 top classes and altogether 176 classes, the immediate research question was how efficiently SVM kernels can reproduce different levels of increasingly fine-grained text categories based on fulltext vs. abstracts only. The corpus was split to 80% training data and 20% test data; validation was not applied. We split the multilabel, multiclass classification problems into one-against-all binary problems and calculated the micro- and macro-averaged precision and recall values, and then their average, the F1 score [35]. For both sets of measurements, this was the most important observation parameter. Only C-SVMs were benchmarked, with the C penalty parameter left at the default value of 1. The implementation used the libsvm library [7]. We used the widespread linear, polynomial and RBF kernels on vectors to study classification performance. Polynomial kernels were benchmarked at second and third degree, RBF kernels were benchmarked with a small value (γ = 1/size of feature set) parameter as well as relatively high one (γ = 1 and 2).

A B-spline kernel with multiple parameters was benchmarked with the length of support ranging between 2 and 10. In terms of the micro- and macroaverage F1 measures, in three out of four cases the wavelet kernel outperformed the traditional kernels while reconstructing existing classification tags based on abstracts (Tables 1 and 2) and full texts (Tables 3 and 4). In all, the wavelet kernel performed best in the task of reconstructing the existing classification on a deeper level from abstracts.

Table 1. Results on top-level categories, abstracts only

Kernel	Linear	Poly	RBF	CSBF				
Support length	-	-	-	2	4	6	8	10
Microaverage P	0.744	0.722	**0.880**	**0.732**	0.713	0.704	0.696	0.695
Macroaverage P	0.685	0.617	**0.976**	**0.685**	0.660	0.657	0.647	0.642
Microaverage R	**0.686**	0.623	0.017	**0.680**	0.672	0.671	0.668	0.666
Macroaverage R	**0.464**	0.398	0.064	**0.469**	0.462	0.467	0.462	0.459
Microaverage F_1	**0.714**	0.669	0.033	**0.705**	0.692	0.687	0.682	0.680
Macroaverage F_1	**0.553**	0.484	0.121	**0.557**	0.544	0.546	0.539	0.536

Table 2. Results on refined categories, abstracts only

Kernel	Linear	Poly	RBF	CSBF				
Support length	-	-	-	2	4	6	8	10
Microaverage P	0.514	0.457	**0.680**	**0.516**	0.503	0.485	0.472	0.466
Macroaverage P	0.603	0.595	**0.951**	**0.611**	0.595	0.572	0.558	0.547
Microaverage R	**0.433**	0.348	0.012	**0.444**	0.441	0.439	0.435	0.433
Macroaverage R	**0.364**	0.295	0.174	**0.362**	0.360	0.361	0.357	0.358
Microaverage F_1	**0.470**	0.395	0.023	**0.478**	0.470	0.461	0.453	0.449
Macroaverage F_1	**0.454**	0.395	0.294	**0.455**	0.449	0.442	0.435	0.432

Table 3. Results on top-level categories, full text

Kernel	Linear	Poly	RBF	CSBF				
Support length	-	-	-	2	4	6	8	10
Microaverage P	0.728	0.591	**0.949**	0.714	0.724	**0.728**	0.712	0.695
Macroaverage P	0.555	0.302	**0.992**	0.490	0.482	**0.538**	0.533	0.532
Microaverage R	**0.738**	0.694	0.568	**0.738**	0.731	0.728	0.713	0.666
Macroaverage R	**0.612**	0.564	0.289	**0.612**	0.588	0.584	0.574	0.459
Microaverage F_1	**0.733**	0.638	0.711	0.726	0.728	**0.728**	0.712	0.680
Macroaverage F_1	**0.582**	0.393	0.448	0.545	0.530	**0.560**	0.553	0.496

Table 4. Results on refined categories, full text

Kernel	Linear	Poly	RBF	CSBF				
Support length	-	-	-	2	4	6	8	10
Microaverage P	0.487	0.335	**0.880**	0.462	0.474	**0.488**	0.480	0.463
Macroaverage P	0.571	0.366	**0.980**	0.556	0.558	**0.573**	0.569	0.557
Microaverage R	**0.523**	0.514	0.230	**0.505**	0.501	0.493	0.485	0.478
Macroaverage R	0.576	**0.586**	0.440	**0.561**	0.499	0.482	0.473	0.463
Microaverage F_1	**0.505**	0.406	0.370	0.482	0.488	**0.491**	0.483	0.471
Macroaverage F_1	0.573	0.451	**0.610**	**0.558**	0.523	0.523	0.521	0.510

9. DISCUSSION

Any kind of evolving content, including semantics, necessarily brings in the concept of phase space to model dynamic systems as a series of consecutive state spaces, where these correspond to stages of a Saltonesque dynamic library [28]. This implication is important for DL in several respects:

- In phase space, a time-dependent system trajectory describes its current state e.g. with regard to semantic or workflow composition. As such trajectories and system states mutually presuppose each other, tracking DL evolution and storing this trail as a new preservable to aid the future reconstruction of past phases of the information lifecycle in an evolving DL clearly calls for new type of metadata;
- Extending the above CSBF kernel to cover phrase and sentence semantics as well can open up the way to

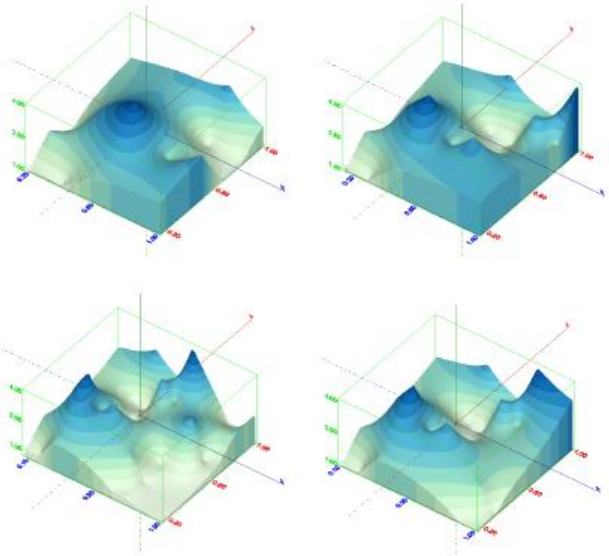


Figure 3. Small-scale model of a changing vocabulary landscape

workflow modeling by two-variable wavelets and thereby a joint coverage of evolving semantic and workflow composition of DLs.

- Flows as the basic components in an information science setting have been identified a long time ago [14]. Provided a semantic arrangement of index terms plus a thematic clustering of documents, another two-variable interpolation method can be used to visualize content dynamics over time or over space. Figure 3 illustrates a changing vocabulary of indexing terminology over time in landscape form: the x and y axes represent the document and term spaces while the z axis the weighted term frequencies, respectively. The individual values are interpolated with appropriate basis functions to emphasize semantic continuity. When studying the spatial propagation of new content, that is, one is to look at the semantic flow at the same time in different geographic regions or in different languages, one sees more differences in the landscape model. On the other hand, there exist methods in TC to address component sequence modeling by wavelets, e.g. for short signals (n-gram models and polynomial kernel up to 2-3 components [8]); shallow parsing and tree kernel [3]; their state-of-the-art combination with bag-of-words (BOW) based semantic smoothing kernels (semantic smoothing tree kernels, [8, 29]); the weighting of sentences [23]; and BOW based multiresolution analysis (MRA) [26].
- Another interesting implication of the Hilbert space approach is that by considering documents as wavelet, i.e. impulse superpositions, one is tempted to assign specific frequencies to index terms and thereby create a hypothetical conceptual spectrum for optical and quantum computing (Figure 4). The bijective mapping that produces this spectrum is based on the semantic ordering of terms, underlying their wavelet representation [13, 33], however, more research will be necessary to bring wavelets and waves on par.

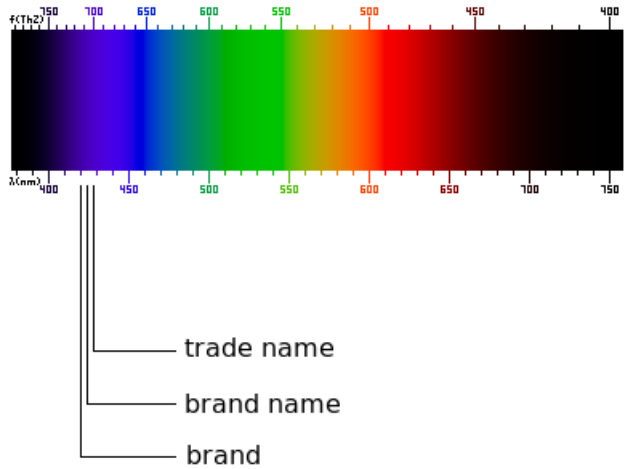


Figure 4. A hypothetical vocabulary spectrum

10. CONCLUSIONS AND FUTURE WORK

A significant impediment to the introduction and use of DLs in developing countries is the lack of appropriate models which address in efficient ways the multicultural and multilingual needs of the users. With the current volumes of DLs such models should promote automated solutions and accommodate methods which facilitate the application of automation. As an answer to this need, we introduced the 5M model of DLs to integrate the management of multicultural, multilingual and multimodal digital objects by multivariate means for advanced access, including the modeling of their evolving semantics in collections. We perceive 5M as a part content-, part form-oriented complex methodological layer built atop of the 5S model of DLs, at the same time connecting to DELOS DLRM.

With all the components of 5M already available in DLs worldwide, what matters is their integration to a next level. To this end, we extended the Spaces component of the 5S model from Euclidean vector spaces to Hilbert spaces. Thanks to this generalization, one can use mathematical objects with higher information representation capacities than vectors, where the difference can be utilized to store different kinds of word semantics and/or word and sentence semantics together. Results from a first experiment with a real-life DL collection, the Strathprints e-repository reflected this semantic surplus. Furthermore, some mathematical objects in Hilbert space, such as vectors and functions, make a non-traditional understanding of word and sentence semantics possible, using evolving physical fields as a metaphor to explain the behavior of language content over time as embodied in, or assigned to, documents. According to this explanation, word content in documents is represented by wavelets whereas sentence and/or document content as sums of their constituent wavelets. This ultimately leads to the conclusion that database content constitutes a phase space.

Researching the viability of the 5M model by benefits of Hilbert space for DLs will have to focus on language representation for the Multilingual dimension as a next step, including the dynamics of language change. We expect sentence semantic kernels to be specific to such spaces, also enabling their application to workflow modeling.

11. ACKNOWLEDGEMENTS

Research for this paper by the first and third authors was funded by the SHAMAN EU project (grant no.: 216736).

12. REFERENCES

[1] Ankerst, M., Breunig, M., Kriegel, H., Sander, J.: OPTICS: Ordering points to identify the clustering structure. In: *Proceedings of SIGMOD-99, International Conference on Management of Data*, ACM Press, New York, NY, USA (1999) 49-60

[2] Baker, A.: Computational approaches to the study of language change. Language and Linguistics Compass 2(3) (2008) 289{307

[3] Bloehdorn, S., Moschitti, A.: Combined syntactic and semantic kernels for text classification. Lecture Notes in Computer Science 4425 (2007) 307

[4] Bohm, D.: Quantum theory. Dover Publications (1989)

[5] Brocks, H., Kranstedt, A., Jäschke, G., Hemmje, M. Modeling Context for Digital Preservation. In Edward Szczerbicki, E., Ngoc Thank Nguyen, eds.: Smart Information and Knowledge Management. Springer, Berlin (2010) 197-226

[6] Candela, L; et al. 2008. The DELOS Digital Library Reference Model - Foundations for Digital Libraries. Version 0.98. Available: http://www.delos.info/files/pdf/ReferenceModel/DELOS_DLReferenceModel_0.98.pdf.

[7] Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. (2001) http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[8] Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press, New York, NY, USA (2000)

[9] Darányi, S., Wittek, P.: On information, meaning, space and geometry. In Turner, P., Davenport, E., Turner, S., eds.: Exploration of Space, Technology and Spatiality: Interdisciplinary Perspectives. Idea Group Publishing, Hershey, PA, USA (2008)

[10] Dawson, A., Slevin, A.: Repository case history: University of Strathclyde Strathprints (2008)

[11] Diaz, F. & Jones, R. (2004).Using Temporal Profiles of Queries for Precision Prediction. *Proceedings of SIGIR'04*, Sheffield, UK (2004) 18-24

[12] Dobreva M., McCulloch E., Birrell D., Feliciati P., Ruthven I., Sykes J., Ünal Y.: Europeana User and Functionality Study: User and Functional Testing. Final report for the EDL Foundation (2010)

[13] Dorrer, C. , Londero, P., Anderson, M., Wallentowitz, S., Walmsley, I.A. Computing with interference: all-optical single-query 50-element database search. *Proceedings of QELS-01, Quantum Electronics and Laser Science Conference* (2001) 149-150

[14] Dretske, F.: Knowledge and the Flow of Information. MIT Press (1981)

[15] Engel, F., Klas, C-P., Brocks, H., Kranstedt, A., Jäschke, G., Hemmje, M. Towards Supporting Context-oriented Information Retrieval in a Scientific Archive-based Information Lifecycle. *Proceedings of Cultural heritage on line. Empowering users: An active role for user communities.* Fondazione Rinascimento Digitale. (2009) Available: http://www.rinascimento-digitale.it/eventi/conference2009/proceedings-2009/engel.pdf

[16] Erk, K.: Representing words as regions in vector space. In: *Proceedings of CoNLL-09, 13th Conference on Computational Natural Language Learning*, Boulder, CO, USA (2009) 57-65.

[17] Eykhof, P. 1974. System Identification: Parameter and State Estimation. Wiley & Sons.

[18] Fellbaum, C.: WordNet: An Electronic Lexical Database. MIT Press, Cambridge, MA, USA (1998)

[19] Goncalves, M., Fox, E., Watson, L., and Kipp, N. 2004. Streams, structures, spaces, scenarios, societies (5s): A formal model for digital libraries. *ACM Transactions on Information Systems* 22(2) (2004) pp. 270-312.

[20] Jiang, J., Conrath, D.: Semantic similarity based on corpus statistics and lexical taxonomy. In: *Proceedings of the International Conference on Research in Computational Linguistics*, Taipei, Taiwan (1997) 19-33

[21] Joachims, T.. Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of ECML-98, 10th European Conference on Machine Learning*, Chemnitz, Gemany (1998) 137-142

[22] Khoo, M., Buchanan, G., and Cunningham**,** S.J. 2009. Lightweight user-friendly evaluation knowledge for digital libraries, *D-Lib Magazine*, July/August 2009, Available: http://www.dlib.org/dlib/july09/khoo/07khoo.html

[23] Ko, Y., Park, J., Seo, J.: Improving text categorization using the importance of sentences. *Information Processing and Management* 40(1) (2004) 65-79

[24] Kumar, V., Bugacov, A., Coutinho, M., and Neches, R. 1999. Integrating geographic information systems, spatial digital libraries and information spaces for conducting humanitarian assistance and disaster relief operations in urban environments. In *Proceedings of the 7th ACM international Symposium on Advances in Geographic information Systems* (Kansas City, Missouri, United States, November 02 - 06, 1999). C. B. Medeiros, Ed. GIS '99. ACM, New York, NY, 146-151.

[25] Lehrer, A.: Semantic fields and lexical structure. American Elsevier, New York, NY, US (1975)

[26] Miller, N., Wong, P., Brewster, M., Foote, H.: TOPIC ISLANDS - a wavelet-based text visualization system. In: *Proceedings of the Conference on Visualization '98, Research* Triangle Park, NC, USA, IEEE Computer Society Press, Los Alamitos,CA, USA (1998) 189-196

[27] Peters, C., Picchi, E. (1997) Across Languages, Across Cultures. Issues in Multilinguality and Digital Libraries D-Lib Magazine, May 1997, ISSN 1082-9873. http://www.dlib.org/dlib/may97/peters/05peters.html

[28] Salton, G.: Dynamic information and library processing. Prentice-Hall, Englewood Cliffs, N.J., US (1975)

[29] Siolas, G., d'Alché Buc, F.: Support vector machines based on a semantic kernel for text categorization. In: *Proceedings of IJCNN-00, IEEE International Joint Conference on Neural Networks*, Austin, TX, USA, IEEE Computer Society Press, Los Alamitos, CA, USA (2000)

[30] Tanner, S. 2010. Technological Trends and Developments and their Future Influence on Digital National Libraries. Appendix A in Hunter D., & Brown, K., Thriving or Surviving? National Library of Scotland in 2030. NLS, 93 pp. Available: http://www.nls.uk/about/policy/docs/future-national-libraries.pdf

[31] Vapnik, V.: Statistical learning theory. John Wiley & Sons, New York, NY, USA (1998)

[32] Weaver, H.: Theory of Discrete and Continuous Fourier Analysis. John Wiley & Sons, New York, NY, USA (1988)

[33] Wittek, P., Darányi, S., Tan, C.: Improving text classification by a sense spectrum approach to term expansion. In: *Proceedings of CoNLL-09, 13th Conference on Computational Natural Language Learning*, Boulder, CO, USA (2009) 183-191

[34] Wittek, P. Compactly Supported Basis Functions as Support Kernels: Capturing Feature Independence in the Embedding Space. PhD thesis. Department of Computer Science, National University of Singapore (2009)

[35] Yang, Y.: An evaluation of statistical approaches to text categorization. Information Retrieval 1(1) (1999) 69-90

[36] Young, N.: An introduction to Hilbert space. Cambridge University Press, New York, NY, USA (1988)