Terminology Based Visualization of Interfaces in Interdisciplinary Research Networks

Thomas Thiele[a], Claudia Jooß[a], Anja Richert[a], Sabina Jeschke[a]

*[a] IMA - Institute of Information Management in Mechanical Engineering &
ZLW - Center for Learning and Knowledge Management &
IfU - Ass. Institute for Management Cybernetics
RWTH Aachen University, Aachen
GERMANY*

Interdisciplinary collaboration between researchers from different fields is often hampered by different wordings and terminologies as well as the varying utilization of similar concepts. Empirical research has shown that a visualization of thematic interfaces between the researchers in a Cluster of Excellence is a possible approach to support scientific collaboration processes.
As one possible solution a technical concept for this approach is outlined in this paper. Terminologies are used to create a visualization of interfaces between the researchers. The paper focuses on the extraction of terminologies and their usage based on scientific publications as well as the corresponding mapping and visualization process using Text Mining algorithms.

Practitioner Summary: A technical concept for a tool is outlined, which supports the collaboration between entities in a (research) network. By creating a visualization based on terminologies the tool enables the user to view the interfaces of his project. This aims at a better efficiency of collaboration in the network.

Keywords: Interdisciplinary Collaboration, Terminologies, Visualization of Interfaces, Text Mining

1. Introduction

Among others the German Joint Science Conference with five central stakeholders[1] jointly stated that cooperation is a key element to a growing success in science (Joint Science Conference, 2009). The processes within these collaborations have to be analyzed, evaluated and fostered in order to determine innovative solutions for complex (scientific) problems, e.g. future production technology or societal challenges like demographic change. The development and usage of synergies between the researchers can be seen as the main aim of this form of collaboration (Jooß et al., 2014). This paper outlines a technical concept to support these collaboration processes and detect synergies, using a statistical approach based on Text Mining algorithms. The visualization of interfaces is derived from publication data and will be implemented as an autonomous self-actualizing tool. The technical concept is described on the basis of an interdisciplinary research network (Cluster of Excellence, CoE) funded by the German DFG. The requirements for this tool result from the structure of the research network on the one hand and from the researchers' needs on the other (cf. Chapter 2.). The main part of this paper contains a technical concept, which aims at visualizing interfaces between entities in the CoE (cf. Chapter 3.).

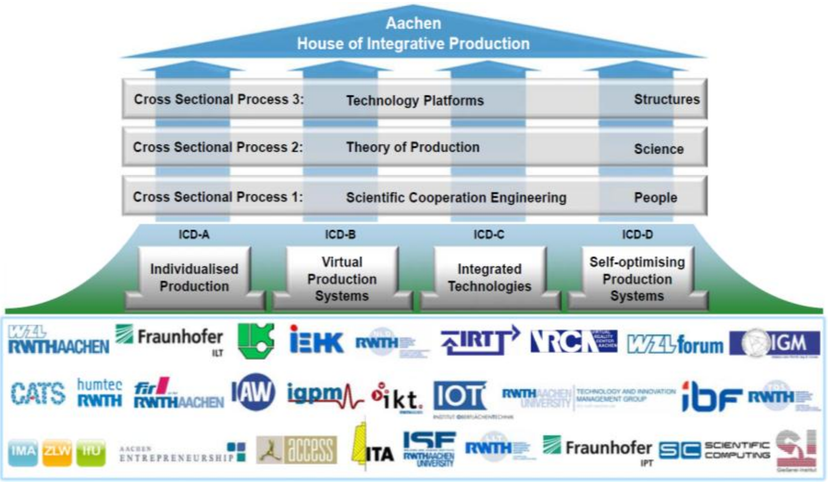
2. Requirements of a Production Technology Research Network

Initiated by the German Research Foundation and the German Council of Science and Humanities the CoE "Integrative Production Technology for High-Wage countries" is part of the German excellence initiative. The cluster is located in Aachen and investigates the solution of the so-called polylemma of production[2]

[1] Namely the Fraunhofer-Gesellschaft, the Helmholtz Association, the Max-Planck Society, the Leibniz Association and the German Research Foundation.

[2] In terms of production technology the polylemma of production depicts the trade-offs between scale and scope on the one hand and between planning and value orientation on the other (Brecher, 2012).

embedded within a holistic theory of production based on integrative comprehension and learning (Brecher, 2012; Cluster of Excellence, 2012) on an interdisciplinary perspective (cf. Figure 1).


Aachen
House of Integrative Production
Cross Sectional Process 3:
Technology Platforms
Structures
Cross Sectional Process 2:
Theory of Production
Science
Cross Sectional Process 1:
Scientific Cooperation Engineering
People
ICD-A
ICD-B
ICD-C
ICD-D
Individualised Production
Virtual Production Systems
Integrated Technologies
Self-optimising Production Systems

Exemplary CoE Figures

32 professors,
180 researchers and
200 student assistants

Interdisciplinary Perspectives

Mechanical Engineering,
Material Science,
Mathematics,
Business Studies,
Communication Science,
Computer Science and
Psychology

Figure 1. CoE structural overview.

This perspective can be seen as the most striking point in order to foster collaboration in the field of production technology: A positive influence on problem solving and knowledge production within interdisciplinary research processes has been determined as main benefits by various sources (Gibbons, 1994; Jacobs & Frickel, 2009; Klein, 1990; Rhoten & Pfirman, 2007; Rijnsoever & Hessels, 2011). To achieve these positive aspects the collaboration of researchers can be supported by various physical and virtual measures (Jooß et al., 2012; Vaegs et al., 2014). These measures focus on an enhanced integration of researchers into the research network (Thiele et al., 2015). All of these efforts are conducted with regard to the idea that especially innovative solutions especially arise on the boundaries of disciplinary borders (Rijnsoever & Hessels, 2011). Therefore a positive impact on innovation capability and processes is assumed, when applying the technical concept as one possible measure in a collaboration process.

Within the context of this paper Jooß demonstrated that a visualization of interfaces as well as the exchange on terminologies are important for the researchers[3]. These can be seen as requirements in the development of collaboration supporting measures:

- “**Exchange and terminology**: A constant communicative exchange is important in the context of interdisciplinary collaboration. This exchange is based on acquiring a common understanding of terms.” (Jooß, 2014, p. 162)
- “**Visualization of the vision**: In the context of interdisciplinary collaboration, it is important to raise awareness for the (interdisciplinary) vision. It is based on the visualization and localization of the individual projects and involved researchers.” (Jooß, 2014, p. 163)
- “**Identification and visualization of interfaces**: A localization of individual researchers involved, projects and contents into the overall context is important as far as interdisciplinary collaboration is concerned. Interfaces can be identified and further processed by means of visualization.” (Jooß, 2014, p. 164)

Following these requirements the technical concept, which is outlined in next chapter, represents one possible outcome in order to support collaboration processes for the researchers within the CoE.

[3] Within this analysis Jooß identified 30 critical incidents (CIs). After having empirically reasoned a theoretical saturation, further and more generally these CIs have been transformed to three patterns (Jooß , 2014, p. 161).

3. Solution approach: Interface visualization

As a possible solution the networking and identification of interfaces can be fostered by a technical concept, which will be implemented as a web application. Based on the usage of terminologies in CoE-publications the concept aims at a visualization in order to depict interfaces and to provide definitions of terminologies to enable the scientists to recognize further research activities in the CoE. Thus, the development and the semantic negotiation of a common understanding regarding terminologies and their usage is supported. At the moment this development mostly takes place at network meetings and workshops. These measures are supported by a technical concept, which the main benefit can be seen in a faster information processing as well as place and time-independent availability of information (Eppler, 2007; Krcmar, 2011).

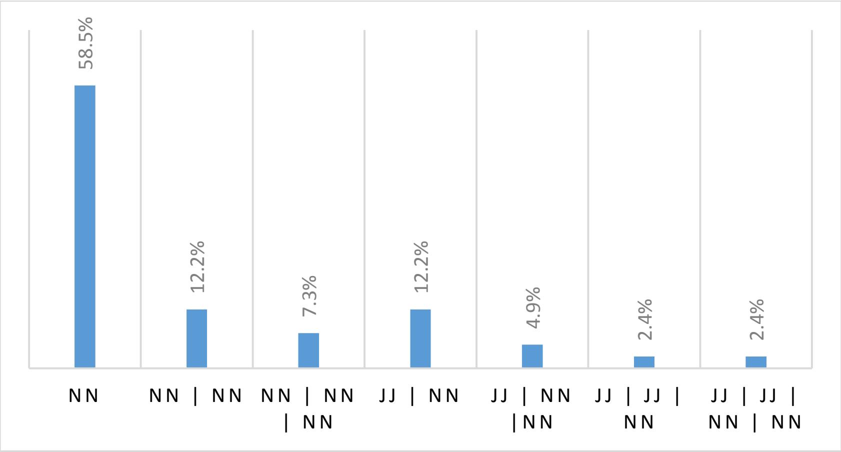
The technical concept derives from the underlying principle of the vector space model. Within this model "[d]ocuments [...] are represented in high-dimensional space, in which each dimension of the space corresponds to a word in the document collection" (Manning & Schütze, 2003). Therefore, the following chapters contain the three major steps which have to be carried out in order to derive a vector based visualization from publications. At first the extraction of terminologies from a CoE-publication database is outlined. In addition, the extraction process contains several preprocessing steps in order to derive a feature vector representation of the publications. The subsequent chapter addresses the mapping of this feature vectors by a combination of classification algorithms and co-occurrence analysis. Since this is the most important part in order to derive a metric for the visualization this paper focuses on this part. In conclusion, chapter 3.3 covers a concept in order to visualize the interfaces and metrics derived in chapter 3.2.

3.1 Extraction of Terminologies

The first step in the identification process of interfaces can be seen in the extraction of terminologies from a given data source. This data source, namely the CoE publications, are accessible via the central CoE web server. Hence, a process is necessary, which transforms the publications from a pdf file into a vectorial form. This so-called feature vector aims at a frequency based description of the publications. This arises two main tasks, which are discussed in this chapter. At first this chapter focuses the process steps, which are necessary in order to derive a feature vector from the publications. Hereafter, the methodology behind the frequency determination is addressed.

As *Text Mining* can be described as "a range of technologies for analyzing and processing semi-structured and unstructured text data" (Miner et al., 2012) this methodology can serve as one possible solution for the challenge mentioned above. According to Miner et al. this range of technology includes processes for the preparation of texts. This includes the tokenization of the publications as well as the filtering of stop words. As one preparation processes for the texts POS-tagging is used to derive nouns from the publications as these word types are considered as main source for terminologies in the CoE[4]. Figure 2 shows the distribution with labels according to the Penn Treebank (Penn Treebank Project, 2003). As NN indicates the usage of nouns in Figure 2 these type of words are used as terminologies in various groupings. With regard to these groupings different combinations of words have to be tagged in order to derive the correct terminologies from the texts, e.g. words composed of adjectives and nouns etc. Due to these compositions various challenges concerning the linguistic properties of the terminologies have to be addressed, e.g. so-called collocations. Collocations can be described as the common habitual occurrence of two or more words (Carstensen et al., 2010). This includes compounds (disk drive) or stock phrases (bacon and eggs) (Manning & Schütze, 2003). With this in mind the given publications are analyzed regarding words that are often used in combinations.

[4] This data is derived from a CoE-internal survey in October 2013. The participants have been invited to send in the terminologies used in their projects.



Exemplary POS-Tags

NN ≙ Noun
JJ ≙ Adjective
| ≙ Logical “and”

Cf. (Penn Treebank Project, 2003)

Figure 2. Distribution of word types within the CoE terminologies.

After that a tf-idf algorithm is applied to determine the frequencies of words within the publications. This algorithm uses a weighting scheme aiming at an effective method to balance term occurrence and document frequency (Manning & Schütze, 2003). In addition, a normalization can be applied. In an experimental test case a common tf-idf variant (Feldman & Sanger, 2007; Manning & Schütze, 2003) using the simple term occurrence $tf_{t,d}$ in combination with logarithmic document frequency df_i has been implemented as follows:

$$w_{i,j} = tf_{t,d} * \log\frac{N}{df_i}$$

In this formula N represents the total number of documents in the collection and $w_{i,j}$ the term weighting in relation to each document *i* and term *j*. In conclusion each calculated frequency is standardized by a cosine normalization. This allows a better comparison between the vector measures as the Euclidean length simplifies to 1. Using a combination of vectors extracted from several publications, which have been issued in the context of a project, a feature vector of an entity (e.g. projects) within the research cluster becomes possible. In the context of this chapter the terminologies are only described by frequency. Although a weighted algorithm has been applied this basic idea can only be an initial step to develop a ranking of importance, which aims at the description of *central* terminologies with regard to interfaces in the CoE. This leads to the next chapter, in which a mapping of entities is outlined.

3.2 Mapping of Entities and Metrics

The second step represents the mapping of entities in order to derive terminology based connections on a statistical approach. By using classification algorithms “the task is to classify a given data instance into a pre-specified set of categories” (Feldman & Sanger, 2007). As the *given data* represents the feature vectors, the labels of the *set of categories* can be seen in the different entities in the CoE, which are described by the combined vectors of several publications (cf. chapter 3.1). Hence, the classification process is implemented in order to elaborate interfaces between the entities (resp. the combined vectors of the entities in the vector space model). Exemplary test results of this process are described in the first part of this chapter, while the second part outlines a co-occurrence based process, which evaluates semantic similarity of vectors by matrix analysis. This constitutes an addition to the frequent analysis implemented by the tf-idf algorithm (cf. chapter 3.1), because a more detailed insight by adding a semantic component to the frequency analysis is given.

In an experimental case a k-nearest-neighbor algorithm has been applied to a test publication set. Within this set all publications have approximately the same length, the training set consists of 7972 (vector) dimensions in three labels. In order to validate the training set it has been cross-validated with a 100% accuracy using $k = 1$. Applying the algorithm with $k = 5$ to the test publication data the result is shown in Table 1. Within this example the test data, publications originated from project four ($P4_{1\ldots5}$), has been classified to the three other labels in this test case (projects 1-3).

Table 1. Excerpt of probability results of experimental test data using a k-nearest-neighbor classifier for k = 5.

Test Data	Label	Probability (P1)	Probability (P2)	Probability (P3)	Prediction
$\boldsymbol{P4_1}$	P4	.200	.400	.400	P2
$\boldsymbol{P4_2}$		.600	.200	.200	P1
$\boldsymbol{P4_3}$			.600	.400	P2
$\boldsymbol{P4_4}$		.400	.200	.400	P1
$\boldsymbol{P4_5}$			.600	.400	P2

The prediction as well as the statistical probabilities can be interpreted as a measure of proximity between the classified entities. The metrics of the visualization will be derived from this classification process. Furthermore, the question arises, how a prediction could be generated, when confidence values are equal (e.g. for $P4_1$the probabilities p_{P1} and p_{P2}). Referring to Bishop ties within algorithm are broken at random (Bishop, 2009), which already leads to the discussion of the test data and the algorithm. Further compensation of the depicted ties in the probabilities will be achieved by extending the training as well as the test data. Although some of the probability values in Table 1 are equal, the fundamental principle becomes obvious. The combination of prediction and probability outlines a statistical connection between two entities (e.g. P1 and P4 on the basis of $P4_2$) and can be used as metric for a visualization. Following this idea the interface between two projects can be described by a vector

$$\overrightarrow{u_k} := f(p_k) \;\; with\, k \;\in \{1; 2\}$$

in this example. The origin of this vector lies in P4 as this is the entity, which has been classified to the other vectors. This example will further be detailed by an exemplary visualization in chapter 3.3.

As outlined at the beginning of chapter 3.2 its second part addresses the semantic similarity of words in order to achieve more detailed insights into the semantics of the terminologies within the CoE. In the context of this paper two non-trivial conditions for a relation between two terms are considered:

- Words may be similar concerning their spelling, but different regarding their meaning.
- Words may be different concerning their spelling, but similar regarding the meaning.

In both cases a partial overlap of meaning may be included in the definitions. This leads to the question on how to derive a threshold for this partial overlap especially for meaning as different spellings can easily be determined via technical means. The acquisition of meaning can be considered as quite a challenge for automated systems. Based on vector space models the focus lies on relative measure for semantic similarity, which can be used to determine how similar a word is to known words by comparing vectors (Fox, 2010; Manning & Schütze, 2003). This aims at the definition of a quantitative similarity measure between different terms. Hence, this measure can be further processed in a visualization.

A first approach to a solution focuses on the analysis of co-occurrence matrixes using the parameter $w_{i,j}$ derived in chapter 3.1. The *latent semantic analysis* (LSA) uses this parameter for the modelling of a similarity matrix following the idea that “two words may be deemed to be conceptually related because the words that they appear with occur together in other documents” (Fox, 2010). In this context the main benefit of LSA is the reuse of already determined data: This approach transforms the weighted frequency of words $w_{i,j}$ to a weighted meaning function m_j that expresses the word’s importance and a measure to which the word is relevant in the current domain of discourse (Landauer, Foltz, & Laham, 1998). Using a dimensionality reduction in the matrix LSA provides the functionality to derive the most important factors $w_{i,j}$. This process results in semantic spaces for most important words, which allows a similarity calculation based on e.g. a cosine similarity, dot product or Euclidean distance (depending on the test case) (Landauer, 2014). This semantic spaces for words can be used, by analogy to $\overrightarrow{u_k}$, to describe a vector as interface between to semantic similar vectors in the form

$$\overrightarrow{t_j} := f\big(\, m_j\big) \;\; with\, j \;\in \mathbb{N}.$$

Following this concept LSA addresses two primary aims in the context of a visualization of interfaces. First, the process allows a semantic reduction focusing only on the most important words in the discourse.

On a second step a quantified similarity between these words is generated. Especially the last step can be seen as a requirement for the visualization of interfaces, which is focused in the next chapter.

3.3 Visualization

The third step in the technical concept represents the visualization of the above mentioned results. The main goal is to depict new interfaces for the researchers on the basis of common terminologies and, therefore, common (research) topics between the user's project and other entities in the CoE. The visualization has to map two major aspects:

- Project to project interfaces derived from the classification process.
- Semantic similarity of words derived from LSA.

On the basis of the examples given in chapter 3.2 a tree graph has been generated (cf. Figure 3). A tree graph allows an easy detection of semantic similarity between words (resp. terminologies, see green lines) and the interfaces between projects (blue lines). Following this idea one goal of visual analytics, the synthesis of information and to communicate this assessment effectively for action is fulfilled (Keim, 2010).

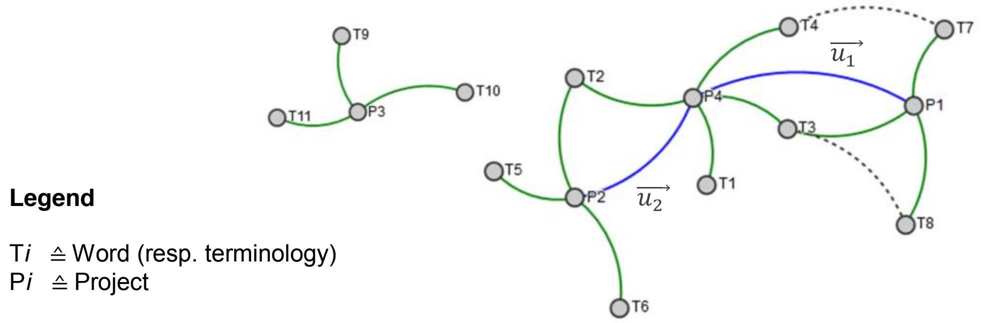


Figure 3. Exemplary visualization of test data.

The exemplary visualization in Figure 3 depicts the different possibilities of the mapping step. The project to project interfaces is based on the example described in chapter 3.2. Centered on P4, the derived vectors $\overrightarrow{u_1}$ and $\overrightarrow{u_2}$ serve as interfaces to P1 and P2. As there is no prediction towards P3 no interface is generated. The magnitude $|\overrightarrow{u_1}|$ and $|\overrightarrow{u_2}|$ are proportionally matched towards the probabilities results of the classification process. Regarding the semantic similarity an exemplary test case showing all possibilities is depicted. P4 and P1 share both a word in spelling and meaning, T3, which is illustrated by a green line. Furthermore, T8 and T3 share the same semantic space, including an overlap of meaning, shown by a dashed line. This example is extended by an interface between P4 and P1: T4, related to P4, shares a semantic space together with T7. All of these interfaces have been generated on the basis of the vectors $\overrightarrow{t_j}$ (not depicted in Figure 3 with regard to legibility), with their length directly proportional to the magnitude $|\overrightarrow{t_j}|$.

4. Conclusion and Outlook

Within this paper a technical concept has been outlined, which fosters collaboration of researchers and enhances the potential for a synergetic collaboration in a research network. The requirement analysis in Chapter 2. showed that the identification and visualization of interfaces as well as the exchange on terminologies (Jooß, 2014) are two major requirements for the technical concept. A possibility for an implementation has been focused in Chapter 3. Within this chapter three major steps (extraction of terminologies, mapping and metrics as well as an exemplary data-driven visualization) are depicted. In summary two levels for the technical concept have been exemplified. On the one hand project to project interfaces are addressed, on the other hand an approach towards word to word interfaces has been revealed.

The next steps can be seen in further validation. Although first test data has been generated and has shown a promising start the main part of the technical proof of concept has yet to be done. Further experiments are also necessary on an algorithmic perspective. With the k-nearest-neighbor classification only a very simple kernel-based method (Bishop, 2009) has been applied in a test case. Further inquiries will continue the proof of concept by a series of experiments for the test publication set. This will include the testing of more complex classification algorithms (e.g. like probabilistic generative models) as well as a cross validation by and testing of clustering algorithms.

Furthermore, a technical concept, although autonomous in most aspects, addresses human needs in the context of a collaboration support. Although good to know, which interfaces exist, the next step has to be a semantic negotiation on a personal level. Hence, the technical concept mainly serves as initiator for further collaboration as well as a display for potential synergies. This idea also offers another perspective with regard to the algorithmic level. Further exchange on acquiring a common understanding of terminologies on a personal level may also be included in the identification process of interfaces in form of a so-called Human-in-the-Loop approach.

Acknowledgements

The authors would like to thank the German Research Foundation DFG for the kind support within the Cluster of Excellence “Integrative Production Technology for High-Wage Countries”.

References

Bishop, C. M. (2009). *Pattern recognition and machine learning. Information science and statistics*. New York, NY: Springer.

Brecher, C. (Ed.). (2012). *Integrative production technology for high-wage countries*. Berlin: Springer.

Carstensen, K.-U., Ebert, C., Ebert, C., Jekat, S. J., Klabunde, R., & Langer, H. (2010). *Computerlinguistik und Sprachtechnologie: Eine Einführung* (3., überarbeitete und erweiterte Auflage). *Spektrum Lehrbuch*. Heidelberg: Spektrum Akademischer Verlag.

Cluster of Excellence. (2012). *‘Integrative Production Technology for High-Wage Countries’: Renewal Proposal for a Cluster of Excellence – Excellence Initiative by the German Federal and State Governments to Promote Science and Research at German Universities*. Unpublished.

Eppler, M. J. (2007). *Managing information quality: Increasing the value of information in knowledge-intensive products and processes* (2nd ed.). Berlin, New York: Springer.

Feldman, R., & Sanger, J. (2007). *The Text Mining handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge, New York: Cambridge University Press.

Fox, C. (2010). Computational Semantics. In A. Clark (Ed.), *A John Wiley & Sons, Ltd., publication. The handbook of computational linguistics and natural language processing* (pp. 394–428). Chichester: Wiley-Blackwell.

Gibbons, M. (1994). *The new production of knowledge: The dynamics of science and research in contemporary societies*. London, Thousand Oaks, Calif.: SAGE Publications.

Hessels, L. K. (2011). Factors associated with disciplinary and interdisciplinary research collaboration. *Research Policy, 40*(3), pp. 463–472.

Jacobs, J. A., & Frickel, S. (2009). Interdisciplinarity: A Critical Assessment. *Annual Review of Sociology, 35,* pp. 43-65.

Joint Science Conference. (2009). *Pakt für Forschung und Innovation: Fortschreibung 2011 – 2015*. Retrieved from http://www.gwk-bonn.de/fileadmin/Papers/PFI-2011-2015.pdf

Jooß, C. (2014). *Gestaltung von Kooperationsprozessen interdisziplinärer Forschungsnetzwerke.* Zugl.: Aachen, Techn. Hochsch, Zugl. Aachen.

Jooß, C., Welter, F., Leisten, I., Richert, A., & Jeschke, S. (2014). Innovationsförderliches Knowledge Engineering in inter- und transdisziplinären Forschungsverbünden. In M. Mai (Ed.), *Handbuch. Handbuch Innovationen* (pp. 105–120). Wiesbaden: Springer VS.

Jooß, C., Welter, F., Leisten, I., Richert, A., Schaar, A.-K., Calero Valdez, A., . . . Jeschke, S. (2012). Scientific Cooperation Engineering in the Cluster of Excellence Integrative Production Technology for High-Wage Countries at RWTH Aachen University. In L. Gómez Chova, A. López Martínez, & I. Candel Torres (Eds.), *ICERI 2012. Conference Proceedings* (pp. 3842–3846). Madrid: International Association of Technology, Education and Development (IATED). Retrieved from http://library.iated.org/view/JOOSS2012SCI

Keim, D. (Ed.). (2010). *Mastering the information age: Solving problems with visual analytics*. Goslar: Eurographics Association.

Klein, J. T. (1990). *Interdisciplinarity: History, theory, and practice*. Detroit: Wayne State University Press.

Krcmar, H. (2011). *Einführung in das Informationsmanagement. Springer-Lehrbuch*. Berlin: Springer.

Landauer, T. K. (2014). LSA as a Theory of Meaning. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis* (pp. 3–34). Psychology Pr.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An Introduction to Latent Semantic Analysis. *Discourse Processes, 25,* 259–284.

Manning, C. D., & Schütze, H. (2003). *Foundations of statistical natural language processing*. Cambridge, Mass.: MIT Press.

Miner, G., Delen, D., Elder, J., Fast, A., Hill, T., & Nisbet, R. A. (2012). The Seven Practice Areas of Text Mining. In G. Miner, D. Delen, J. Elder, A. Fast, T. Hill, & R. A. Nisbet (Eds.), *Practical text mining and statistical analysis for non-structured text data applications* (pp. 29–41). Amsterdam: Elsevier/Academic Press.

Penn Treebank Project. (2003). *Alphabetical list of part-of-speech tags used in the Penn Treebank Project*. Retrieved from https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

Rhoten, D., & Pfirman, S. (2007). Women in interdisciplinary science: Exploring preferences and consequences. *Research Policy, 36*(1), pp. 56–75.

Rijnsoever, F. J., & Hessels, L. K. (2011). Factors associated with disciplinary and interdisciplinary research collaboration. *Research Policy, 40*(3), 463–472.

Thiele, T., Schröder, S., Calero Valdez, A., Jooß, C., Richert, A., Ziefle, M., . . . Jeschke, S. (2015). Unterstützung interdisziplinärer Integration am Beispiel einer Exzellenzcluster-Community. In G. Schuh, V. Stich, E.-M. Jakobs, & M. Ziefle (Eds.), *FIR-Edition Forschung Band 15. Zukunft gestalten. Soziale Technologien in Organisationen in Zeiten des demografischen Wandels* (pp. 205–213). Aachen: AWD Druck + Verlag GmbH.

Vaegs, T., Calero Valdez, A., Schaar, A.-K., Bräkling, A., Aghassi, S., Jansen, U., . . . Jeschke, S. (2014). Enhancing Scientific Cooperation of an Interdisciplinary Cluster of Excellence via a Scientific Cooperation Portal. In David Guralnick (Ed.), *Proceedings of the Seventh International Conference on E-Learning in the Workplace* . New York, NY, USA. Retrieved from http://www.icelw.org/program/ICELW%202014%20Proceedings/ICELW2014/papers/Vaegs_et_al.pdf