Der Historical BioData Explorer: Historische Texte, Bilder und Objekte neu erforschen

Michael Dürfeld[a], Philipp Schröter[a], Anika Schultz[a]

Michaela Eder[b], Christian Stein[a], Friederike Saxe[a], Clemens Schmitt[b], Markus Mandalka[a], Svantje Lilienthal[a], Wolfgang Schäffner[a], Peter Fratzl[b]

[a] Interdisziplinäres Labor Bild Wissen Gestaltung, Humboldt-Universität zu Berlin, Deutschland, michael.duerfeld@hu-berlin.de; [b]Abteilung für Biomaterialien, Max-Planck-Institut für Kolloid- und Grenzflächenforschung Potsdam, Deutschland, Michaela.Eder@mpikg.mpg.de

KURZDARSTELLUNG: Der *Historical BioData Explorer* (HBDX) macht historische, biologische Forschungsdaten zu Bewegungen im Tierreich in ihrer medialen Breite von digitalisierten Texten, Bildern und Objekten über ein integratives, bildorientiertes Web-Interface durchsuchbar. Grundlage dafür ist eine eigens entwickelte ontologiebasierte Datenbank, welche das historische Material für aktuelle Fragestellungen aus der (bioinspirierten) Material- und Ingenieurwissenschaft, Architektur und Kultur- und Bildwissenschaft zugänglich macht.

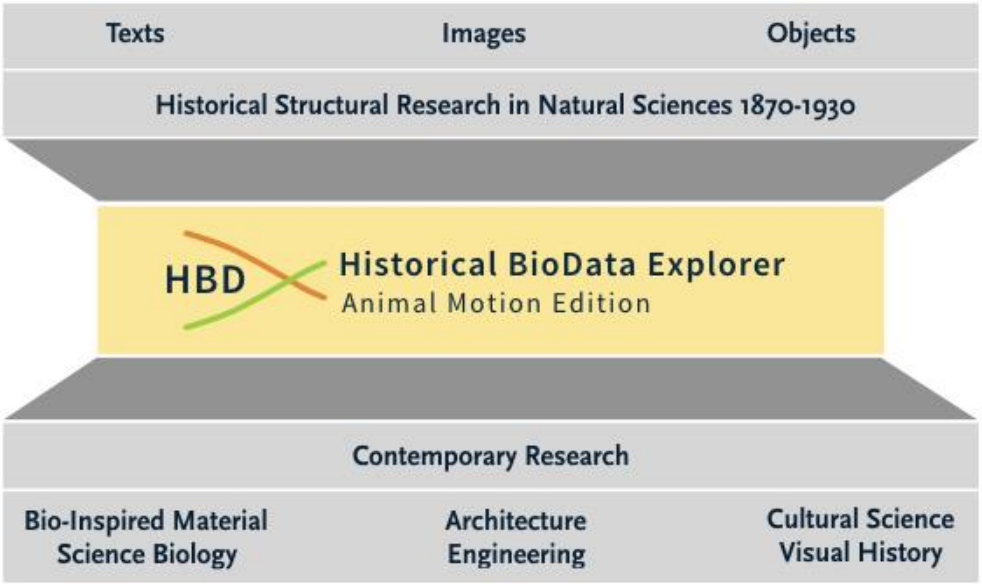

Abb.1: ***PROJEKTKONZEPT DES HISTORICAL BIODATA EXPLORER***

1. HISTORISCHE STRUKTUR-UNTERSUCHUNGEN IM LABOR

Der HBDX ist Teil des Forschungsprojektes "Historische Strukturuntersuchungen im Labor" im Interdisziplinären Labor *Bild Wissen Gestaltung* an der Humboldt-Universität zu Berlin. Im Interdisziplinären Labor erforschen Wissenschaftler_innen aus über 40 Disziplinen grundlegende Gestaltungsprozesse in den Wissenschaften. An dem Zusammenschluss aus Geistes-, Natur- und Technikwissenschaften und der Medizin sind auch die Gestaltungsdisziplinen Design und Architektur beteiligt.

Ausgangspunkt für das interdisziplinäre Team des Forschungsprojektes "Historische Strukturuntersuchungen im Labor" aus Biologie, Kulturwissenschaft, Architektur, Design und Informatik ist die Tatsache, dass die vergleichende Beschreibung von Anatomie und Morphologie einer Vielzahl von Organismen zwischen etwa 1870 und 1930 auf hohem Niveau betrieben und die Ergebnisse sehr detailliert in Text und Bild festgehalten wurden [1][2]. Mitte des 20. Jahrhunderts nahm die Bedeutung der Disziplinen Genetik und Molekularbiologie weitgehend zu und die zeitintensive, detaillierte Erforschung biologischer Systeme konzentriert sich bis heute auf wenige Modellorganismen. Die historischen Forschungsergebnisse zu einer Vielzahl von Organismen in Form von Texten, Bildern und Objekten sind daher gefährdet, in Vergessenheit zu geraten.

Unabhängig davon ist eine Weiterentwicklung der Strukturenuntersuchungen von D'Arcy Thompson zu Buckminster Fuller erfolgt und hat von dort Eingang in die Gestaltungspraxis von Architektur, Städtebau und Ingenieurwesen (Christopher Alexander, Peter Eisenman, Norman Foster) gefunden. Die kulturwissenschaftliche Forschung hat vor dem Hintergrund dieser vielfältigen Strukturuntersuchungen sowohl der Naturwissenschaften als auch der Gestaltungsdisziplinen in den letzten Jahren begonnen, jenseits des rein geisteswissenschaftlichen Strukturalismus eine transdisziplinäre Strukturengeschichte zu entwickeln [3]. Unser Projekt verbindet diese unterschiedlichen Perspektiven und entwickelt einen material-, struktur- und funktionsspezifischen Explorer, der als interdisziplinäres Tool die Auffindung von vielfältigen Strukturen und Formen (in der Form von Bildern, Objekten und Texten) für Fragestellungen aus dem Design, der Architektur oder der Materialerforschung und –entwicklung ermöglichen soll.

1.1 Die historischen Materialien

Das Projekt reaktiviert relevantes historisches Wissen über morphologische Struktur-Untersuchungen aus dem Zeitraum 1870-1930. Bei der Neuerschließung fokussierten wir die Arbeit zunächst auf Texte und Zeichnungen des "Handbuchs der Zoologie" [4] und auf die "Berichte der Challenger Expedition" [5]. Das *Handbuch der Zoologie beschreibt* seit 1923 bis heute fortlaufend die Gesamtheit der bekannten Organismen in detailierten Texten und Bildern - meist in Form von Strichzeichnungen. Mit den *Berichten der Challenger Expedition* von 1880-95 werden weitere Texte und Lithographien eingebunden. Die Berichte sind die Ergebnisse einer Expedition (1872-76) zur Untersuchung der geologischen und zoologischen Beschaffenheit der Ozeanböden und stellt damit den Anfang der modernen Ozeanografie dar. In den 50 Bänden sind zum Teil Lithografien enthalten von höchster Detaillierung und gestalterischen Qualität - zu den bekanntesten zählen Lithographien von Ernst Haeckel. Zusätzlich wurde die Einbindung historischer Objekte in Form von Präparaten und Modellen aus den naturwissenschaftlichen Sammlungen der Humboldt-Universität vorgenommen; insbesondere historische Präparate und Modelle aus der digitalen Sammlung “Kabinette des Wissens”[6].

1.2 Aktueller Stand der digitalen Wissensspeicherung

Für Materialwissenschaftler_innen interessante Datenbanken aus dem biologischen Bereich konzentrieren sich meist auf aktuelle Forschungsdaten - historisches Material insbesondere digitalisierte Bilder oder Objekte hingegen finden sich dort kaum. Auch geht das hier gespeicherte aktuelle Forschungswissen eher in die Tiefe als in die Breite. Die Erforschung einer Vielzahl biologischer Organismen ist durch die zunehmende Spezialisierung in der Biologie in den Hintergrund getreten. Die hier entwickelte Datenbank und die dem HBDX zugrundeliegende *AnimalMotionOntology*

haben einen anderen Ansatz als aktuelle Projekte in der natur- und material-wissenschaftlichen Forschung [7]-[11]: Es wird auf auf maximale Genauigkeit und Tiefe verzichtet, um das Wissen über Bewegungen im Tierreich so zu modellieren, dass es allgemeinverständlich erfasst werden kann.

Ein weiteres Merkmal ist, dass historisches Material sich zur Zeit durchaus auch in anderen Datenbanken [12][13] findet - diese sind jedoch nach bibliothekarischen Kriterien durchsuchbar. Zudem macht der HBDX eine bildorientierte Suche stark und versucht so, die medial unterschiedlichen Wissensträger zu vernetzen. Erst dadurch wird es möglich, dass das Wissen für mehrere Disziplinen verständlich und greifbar wird.

2. HERAUSFORDERUNGEN, STRATEGIEN UND WERKZEUGE

Die Herausforderungen des Projektes bestehen u.a. darin, bildfixiertes Wissen in textuelles, durch Strukturierung und Semantisierung maschinenlesbares Wissen zu transformieren, disziplinspezifische Wissensordnungssysteme und Terminologien zu verbinden und medienspezifische Wissensrepräsentationen herauszuarbeiten.

2.1 Wissensmodellierung

Um das in Texten, Bildern und Objekten gespeicherte historische Wissen aus den analogen Medien (1) adäquat zu erschließen, (2) neu im digitalen Medium darzustellen und (3) maschinenlesbar aufzubereiten, muss zunächst eine digitale Wissensmodellierung gewählt und durchgeführt werden. Im Forschungsprojekt wird eine ontologiebasierte Wissensmodellierung verwendet, da - im Unterschied zu einer Taxonomie, die nur eine hierarchische Untergliederung bildet - eine Ontologie die Informationen in einem Netzwerk mit logischen Relationen abbildet. Diese logischen Relationen ermöglichen es, Rückschlüsse aus vorhanden Daten zu ziehen, die vorher nicht beobachtbar waren.

2.2 Ordnungssystematik

Die historischen Daten sind nach biologischen Kriterien und Systematiken organisiert und damit für andere Disziplinen in der Originalform teilweise schwer zugänglich. Deshalb wurden die Daten aus dem biologischen Ordnungssystem herausgelöst und in ein neues System überführt, das einen einfacheren disziplinübergreifenden Zugriff ermöglicht. Die im Forschungsprojekt dazu entwickelte Ontologie kann als eine "Brücken-Ontologie" verstanden werden, die ausgehend von dem historischen Wissenspool, sowohl aktuelle biologisch-morphologische als auch material- und ingenieurwissenschaftliche und bild- und kulturwissenschaftliche Ontologien verknüpft.

2.3 Medientransformation

Die historische Literatur besteht mit ihren Texten und Bildern aus unterschiedlichen Medien, die jeweils unterschiedlich Wissen gestalten, d.h. auswählen und darstellen. Deshalb wurden unterschiedliche Strategien entwickelt, um das medienspezifisch gestaltete Wissen in die digitale Wissensrepräsentation der Ontologie zu überführen: Während Textinformationen mit Hilfe von Text Mining Software automatisiert erschlossen und verarbeitet werden können, müssen Bildinformationen einzeln manuell annotiert werden. Zukünftig besteht hier die Möglichkeit, den Einsatz von automatisierter Bilderkennung durch Mustererkennungs-software zu erproben.

2.4 Terminologie

Die Texte der Materialbasis sind in der Fachsprache der Biologie formuliert, während die Suchanfragen in material-wissenschaftlichen, ingenieurtechnischen und kulturwissenschaftlichen Fachsprachen formuliert werden. Hier muss ein *Matching* der unterschiedlichen Fachsprachen und Fachtermini durchgeführt werden. Dazu werden einerseits Thesauri zusammengestellt, die biologische Begriffe mit materialwissenschaftlichen Begriffen zusammen führen und andererseits sollen Techniken der Natural Language Analysis erprobt werden.

2.5 Sprach-Dualität

Die historischen Texte sind teilweise auf Deutsch geschrieben, während die aktuelle Materialforschung meist auf Englisch kommuniziert wird. Deshalb müssen englisch formulierten Suchbegriffe übersetzt werden und auf die entsprechenden deutschen Begriffe im historischen Korpus verweisen. Hierfür werden digitale Wörterbücher und spezielle Dienste verwendet, die ein solches *Sprachmapping* durchführen.

2.6 Historizität

Die Historizität der Materialbasis bedingt, dass die Inhalte und deren mediale Darstellungen in Sprache und Bild durch den historischen Erkenntnisstand, das historische Erkenntnisinteresse, die historische Sprachverwendung und die historischen Bild-Darstellungstechniken geprägt sind.

Dadurch ergeben sich Differenzen zur aktuellen Forschung. Diese werden herausgearbeitet, um zu erkennen, wo das historische Wissen Innovationspotential besitzt, indem es noch nicht beachtete Sachverhalte zeigt und wo es mit dem aktuellen materialwissenschaftlichen und biologischen Wissen ergänzt werden muss.

3. DIE FUNKTIONSELEMENTE DES HBDX

Diese Grundlagenforschung wird in den konkreten ineinandergreifenden Funktionselementen des HBDX angewendet: In der *AnimalMotionOntology*, in der *Historical BioData Search Engine*, im Bildannotationswerkzeug *TripleGeany* und im Interface Design.

3.1 AnimalMotionOntology

Der zentrale Baustein des HBDX ist die AnimalMotionOntology, mit der das Wissen über Bewegungen im Tierreich modelliert wird. Hier werden alle für Bewegungen relevanten Faktoren sprachlich gefasst, formal und logisch strukturiert und maschinenlesbar codiert.

Dafür wurde herausgearbeitet, welche verschiedenen Bewegungsformen (laufen, hüpfen, schlängeln etc.) unterschieden werden können, in welchen Medien (Wasser, Erde, Luft, Holz etc.) sich Organismen bewegen, welche Elemente (Flügel, Beine, Muskeln, Sedimente etc.) an der Bewegung beteiligt sind, aus welchem Material (Protein, Zucker, Mineral etc.) die Elemente bestehen, welche Materialeigenschaften (fest, flüssig, elastisch

etc.) diese Materialien aufweisen und welche Strukturen (Symmetrien, Muster, Geometrien etc.) die Elemente bzw. der ganze Organismus aufweisen. Ergänzt werden diese Daten durch

Abb.2: Funktionselemente des HBDX

Metadaten, die festhalten, wo (auf welchem Bild, auf welcher Seite, in welchem Band etc.) und wie (als Foto oder Zeichnung, als Perspektive, Schnitt oder Ansicht etc.) der Organismus abgebildet ist. Die formale Ordnung wird durch die Formulierung von logischen Aussagen über die einzelnen herausgearbeiteten Faktoren durchgeführt. Dazu wird das sogenannte Resource Description Framework (RDF) [14] verwendet. Im RDF-Modell besteht jede Aussage aus den drei Einheiten Subjekt, Prädikat und Objekt („Tripel"), wobei das Subjekt durch ein semantisch qualifizierendes Prädikat mit einem Objekt verbunden wird. Ebenso können mit dem RDF-Modell Hierarchien und Klassenzuordnungen formalisiert dargestellt werden. Jedes Tripel stellt eine logische Aussage dar, und kann wiederum mit anderen Tripeln logisch verknüpft werden. Auf der Grundlage dieser Ontologie können nun die Bilder aus der historischen Materialbasis annotiert werden. Jedes Bild wird dazu sprachlich gefasst und die darin befindlichen Informationen in Form von Tripeln formalisiert.

3.2 OCR

Die historischen Werke lagen ursprünglich nicht in einem digitalen Format vor. Um die Texte in späteren Schritten weiterverwenden zu können, wurde während des Scanvorgangs Optical Character Recognition (OCR) [15] angewendet. Damit liegen die Texte vollständig digital und maschinenverarbeitbar vor.

3.3 HBDS - Historical BioData Search Engine

Bevor jedoch Bilder auf der Grundlage der AnimalMotionOntology annotiert werden können, müssen zunächst annotationsrelevante Bilder gefunden werden, d.h. Bilder, die Bewegungen abbilden. Da sich allein in dem

Handbuch der Zoologie über 30.000 Bilder befinden, wurde eine eigene Suchmaschine durch das Projektmitglied Markus Mandalka aufgesetzt (Apache Solr) [16]. Diese indexiert alle Texte, kann Bildunterschriften von normalem Fließtext unterscheiden und diese Bildunterschriften den jeweiligen Bildern

zuordnen. Dadurch wird eine Suche nach auf der Ontologie basierenden bewegungs-anzeigenden Wörtern ausschließlich in den Bildunterschriften ermöglicht. Das Ergebnis ist eine Liste von Bildern, mit dem in der Bildunterschrift enthaltenen Suchbegriff und der entsprechenden Stelle im Digitalisat. Für eine effektive Suche wird dann auf den Index zurückgegriffen, der ebenfalls konjugierte Formen der Bewegungsbegriffe beherrscht. Desweiteren werden die Worthäufigkeiten analysiert und in verschiedenen Varianten aufbereitet, z. B. nach Kapiteln und dem gemeinsamen Vorkommen untereinander. Diese Listen können neben einer einfachen Wortsuche auch als Ausgangspunkt für eine Suche oder als Filter eines Suchergebnisses verwendet werden.

3.4 Bildannotationswerkzeug *TripleGeany*

Für die manuelle Annotation der relevanten Bilder wurde die Webanwendung *TripleGeany*

von der Projektmitarbeiterin Svantje Lilienthal im Rahmen einer Masterarbeit entwickelt. Das

formulargestützte Eingabetool übersetzt die eingegebenen Daten automatisch in RDF-Tripel und speichert sie in einem *Triplestore* ab. Die einzelnen Module des Formulars können durch einen Administrator frei nach Bedarf der entsprechenden Ontologie konfiguriert werden. Die Annotierenden hingegen brauchen kein Wissen über die spezifische RDF-Struktur oder die Ontologie - sie können sich ganz darauf konzentrieren, was auf den Bildern zu sehen ist und dafür die entsprechenden Felder im Formular auszufüllen - meist durch anwählbare vorkonfigurierte Menüs. Am Ende eines solchen Annotationsvorgangs ist die bewegungsrelevante Information auf dem Bild in ein sprachliches und maschinenlesbares Format transformiert und gespeichert worden. Neben der manuellen Annotation werden auch Methoden der Named Entity Extraction und des Named Entity Linking erprobt. Dadurch soll die Datenmenge erheblich vergrößert werden. Dafür werden die bei der Indexierung des Textes verwendeten Informationen genutzt, um daraus automatisiert Tripel mit Informationen zu generieren.

3.5 Interface Design

Über das Interface können Nutzer_innen das historische biologische Material erkunden. Bei dem Entwurf des Explorers ist unser Anspruch gewesen, einen disziplinübergreifenden, visuellen und unmittelbaren Zugang zu dem

historischen Material zu gestalten, welches kein Fachwissen voraussetzt. Deshalb wurde die Startseite so gestaltet, dass Nutzer_innen mit wechselnden Suchvorschlägen zu Bewegungsarten und entsprechenden Bildbeispielen empfangen werden. Dies soll Interesse für die vielfältig illustrierten

des Inhalts im Originalwerk. Darüberhinaus soll es Verlinkungen des historischen Materials zu anderen Forschungsdatenbanken. Neben dieser Sektion gibt es noch eine allgemeine Übersichtsseite, welche das Projekt und den HBDX vorstellt. Alles in allem ist der Funktionsumfang des Explorers überschaubar

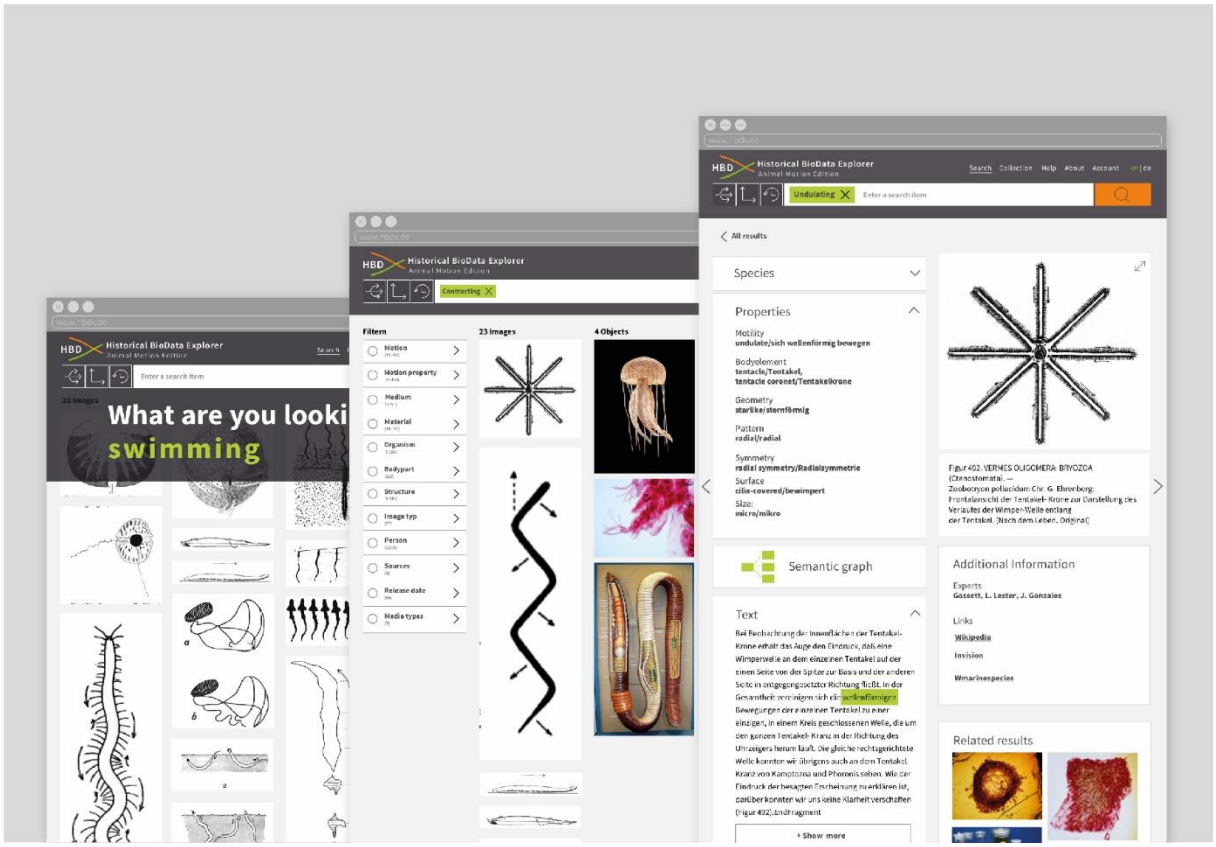

Abb.3: Screendesigns des Interfaces

Bewegungsuntersuchungen wecken und Unerwartetes aufzeigen. Neben den zur Anregung dienenden Vorschlägen, können Nutzer_innen auch direkte Suchanfragen stellen oder die auf die Ontologie basierenden Filter nutzen. So können, alle an der Bewegung beteiligten Faktoren erkundet oder sich medienspezifisch angezeigt werden. Ein weiterer gestalterischer Anspruch war es, ein müheloses Ein- und Auszoomen in das historische Material zu ermöglichen und umstandslos zwischen Übersicht und Detailansicht wechseln zu können. Auf der Übersicht sind die Informationen medienspezifisch sortiert und bei Mouseover steckbriefartig zusammen-gefasst. Auf der Detailseite befinden sich Informationen zum Organismus (Name und taxonomische Einordnung), den im Bild dargestellten Bewegungseigenschaften sowie die Verortung

und konzentriert sich darauf, die verknüpften Inhalte übersichtlich, in ansprechender Weise und medienadäquat darzustellen.

4. DANKSAGUNG

Wir danken unseren studentischen Mitarbeiter_innen Linda Winkler und Lisa O`Connor für die Unterstützung zu Beginn des Projektes, Anne Lange für die Unterstützung im Design und Aydan Cakir, Charlene Faustin, Eliana Campos Zapata, Kilian Weil and Paulina Nowak für die OCR und Bildannotation. Für die interessante Diskussion danken wir Yves Brechet, Lars Vogt, Peter Grobe, Gerhard Scholtz, Joachim Krausse und James Weaver - ihm danken wir auch für die leihweise Überlassung der Original Challenger-Reports zum Scannen. Hierbei danken wir Viola Rosenau vom *Kulturgutscanner* für deren exzellente Arbeit

in der Digitalisierung. Jochen Henning danken wir für den Zugriff auf die Daten der *Kabinette des Wissens*. Das Projekt ist ein Projekt des Exzellenzclusters Bild Wissen Gestaltung. Ein Interdisziplinäres Labor der Humboldt-Universität zu Berlin, gefördert durch die Deutsche Forschungsgesellschaft.

5. LITERATURHINWEIS

[1] Fratzl, P.; Weiner, S. Bio-Inspired Materials–Mining the Old Literature for New Ideas, *Advanced Materials*. 2010, *22*, 4547-4550 .

[2] Coleman, W. Biology in the nineteenth century: problems of form, function and transformation: Cambridge University Press, 1971.

[3] Schäffner, Wolfgang (2008): Ein neuer Strukturalismus – zur Gestaltung des Wissens in einem interdisziplinären Strukturen-Labor. In: Krausse, Joachim/Pinkau, Stephan (Hg.): The Intelligence of Structures. Bauhaus Lectures Dessau. Dessau: Stiftung Bauhaus u. a., S. 118–129.

[4] Kükenthal, W. G.; Krumbach, T. Handbook of zoology: a natural history of the phyla of the animal kingdom: Walter de Gruyter, 1923.

[5] Report on the scientific results of the voyage of H.M.S. Challenger during the years 1873-76 under the command of Captain George S. Nares ... and the late Captain Frank Tourle Thomson, R.N.

[6] http://www.sammlungen.hu-berlin.de/kdw (Zugriff am 20.10.2017)

[7] Deldin, J.-M.; Schuknecht, M. The AskNature Database: enabling solutions in biomimetic design. In *Biologically inspired design*). Springer, 2014, 17-27.

[8] Chakrabarti, A.; Sarkar, P.; Leelavathamma, B.; Nataraju, B. A functional representation for aiding biomimetic and artificial inspiration of new ideas, *AIE EDAM*. 2005, *19*, 113-132.

[9] Vincent, J. F. V.; Bogatyreva, O.; Pahl, A. K.; Bogatyrev, N.; Bowyer, A. Putting biology into TRIZ: a database of biological effects, *Creativity and Innovation Management*. 2005, *14*, 66-72.

[10] Vincent, J. F. V. An Ontology of Biomimetics. In *Biologically Inspired Design: Computational Methods and Tools* (Goel, K. A., McAdams, A. D. and Stone, B. R. (eds.)). London: Springer London, 2014, 269-285.

[11] Grobe, P.; Vogt, L. MorphDBase-A Morphological Description Database, *Proccedings of Conference MorphDBase-A Morphological Description Database*, 2008, *269*, 1478-1479.

[12] https://www.biodiversitylibrary.org/ (Zugriff am 20.10.2017)

[13] https://www.europeana.eu/portal/en

[14] (Zugriff am 20.10.2017)

[15] https://www.w3.org/RDF/

[16] (Zugriff am 20.10.2017)

[17] Verwendetes Programm:

[18] Abbyy FineReader Pro

[19] http://lucene.apache.org/solr/