2014

Semantic Blumenbach

Linking Structured Text and Structured Data

The following paper describes the independent study project of Christopher H. Johnson conducted from 6 September 2013 to 01 February 2014.

The study project examined the initiative “Semantic Blumenbach” developed by the Academy of Science and Humanities at Göttingen in a project of the Lower Saxony Digital Humanities Research Collaboration (DHFV) at the Göttingen Centre for Digital Humanities (GCDH) to explore and apply Semantic Web technologies to establish methods for providing and presenting linked data. These technologies model the semantic relationships between objects described by TEI-encoded texts of Johann Friedrich Blumenbach and metadata of these items stored today in several University collections

Christopher H. Johnson
#3147338
Brandenburg Technische Universität Cottbus-Senftenberg
Submitted to: Professor Michael Schmidt
22-Feb-14

CONTENTS

GLOSSARY OF ACRONYMS

TERM	DESCRIPTION
ABOX	Assertion Component of an Ontological Axiom
BTU	Brandenburg Technische Universität
CERL	Consortium of European Research Libraries
CIDOC CRM	Cultural Information Documentation Conceptual Reference Model
DFN	Deutsches Forschungsnetz
DL	Description Logic
EAD	Encoded Archival Description
GATE	General Architecture for Text Engineering
GCDH	Göttingen Center for Digital Humanities
GND	Gemeinsame Normdatei
LOD	Linked Open Data
NER	Named Entity Recognition
NLP	Natural Language Processing
OWL	Web Ontology Language
PHP	Hypertext Preprocessor General Purpose Scripting Language
RDF	Resource Description Framework
RNG	Regular Language for XML Next Generation
SLUB	Staats- und Universitätsbibliothek Dresden
SPARQL	SPARQL Protocol and RDF Query Language
SWRL	Semantic Web Rule Language
TBOX	Terminological Component of an Ontological Axiom
TEI	Text Encoding Initiative
TXM	Textometrie
URI	Uniform Resource Identifier
WissKI	Wissenschaftliche Kommunikations Infrastruktur
XML	Extensible Markup Language
XSLT	Extensible Stylesheet Language Transformations

1. INTRODUCTION

The main purpose of the Semantic Web is to enable users to find, share, and combine information more easily. The Semantic Web is a system that enables machines to "understand" the content, links and transactions of data based on their meaning. For the machines to understand and analyze this data requires that the relevant information sources be semantically structured. The semantic structuring of historic information sources is the focus of this paper.

The project "Semantic Blumenbach" defines its objective as "to discover and render visible the intense connections of Blumenbach's writings with the objects studied by him" (GCDH, 2014). Johann Friedrich Blumenbach (11 May 1752 – 22 January 1840) was a German anatomist, zoologist and anthropologist. He is regarded as a key founder of zoology and anthropology as scientific disciplines. Blumenbach was a prolific writer as well as an avid collector of objects, ranging from skulls and bones to rocks and minerals. The digitization and creation of the collection object repository has been the focus of the parent project of Semantic Blumenbach, "Blumenbach Online".

The project "Blumenbach Online" is in total defined with the following objectives (SUB, 2014):

1. A new edition of Blumenbach's original works including their translations and reissues.
2. An annotated calendar of Blumenbach's correspondence.
3. An inventory and reconstruction of Blumenbach's scientific collections.
4. A documentation of his contemporary and later reception.
5. The hyperlinking of the digital texts and objects.
6. Biographical studies of Blumenbach.

Objective 5 is therefore the main task of Semantic Blumenbach. The task of "hyperlinking", while it sounds relatively simple, is comprised of a diverse array of advanced technical tools and methods that will be described in detail in the following sections.

2. PROJECT STRUCTURE AND DEFINITION

The Semantic Blumenbach Project has been designed with a three tier structure consisting of 1) Named Entity Recognition (NER) 2) Translation and 3) Extension illustrated in Figure 1. The Named Entity Recognition components have recently been completed, and the methods and concepts of NER have been a dominant part of this study project's focus. The Tier 2 Translation and Tier 3 extension components have also been evaluated within the scope of this study project, but the specific method development remains a work in progress.

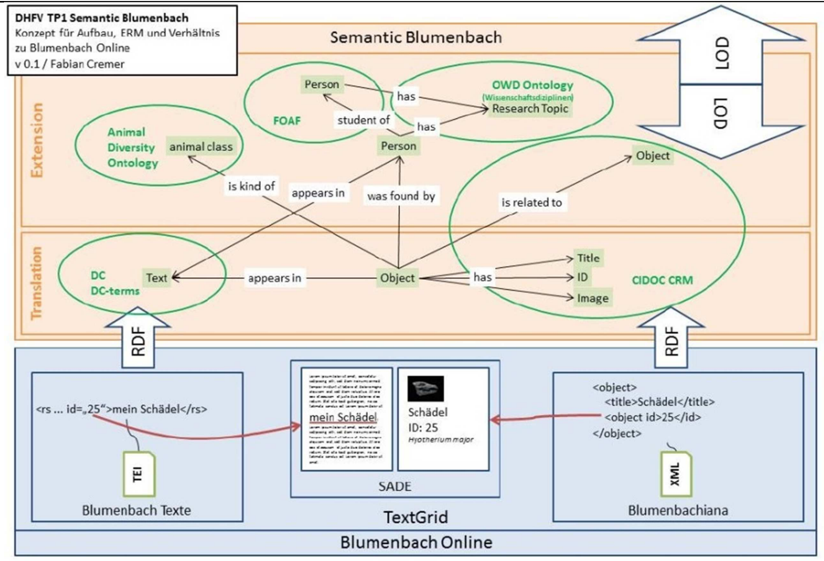


Figure 1 Semantic Blumenbach Structure

WISSKI SYSTEM
EXTENSION

CIDOC CRM
TRANSLATION

TEI
NAMED ENTITY RECOGNITION

PROJECT OBJECTIVES

1. Provide proof of concept for Semantic Web Modelling of relationships between Blumenbach's texts and collection items in Blumenbach Online.
2. Publication of Linked Open Data
3. Developing generic tools for projects in the Academy of Science

The Project will be finished by End of March 2015. Evaluation will be take place in autumn 2014.

TIER 1: NAMED ENTITY RECOGNITION

Named-entity recognition (NER) is a subtask of information extraction that seeks to locate, annotate and classify elements in a text into predefined categories such as persons, places, technical terms, objects and dates. A particular feature of the texts written by Blumenbach is the highly specialized scientific vocabulary originating from the field of natural history and including vernacular idioms that are out of use today. Rather than starting from ordinary unstructured text, a specialized type of structured text, known as TEI, has been used for input for the GCDH NER process.

TEI: TEXT ENCODING INITIATIVE

Since 1987, The Text Encoding Initiative (TEI-C) is a consortium which collectively develops and maintains guidelines for encoding machine-readable texts in the humanities and social sciences. The established source platform for Blumenbach TEI texts is the TextGrid Repository (http://www.textgridrep.de/). The project has currently selected one particular encoded text, the 1799 *Handbuch der Naturgeschichte* of Blumenbach, for experimentation of NER. This file is identified generally in short as "000027.xml" based on the coded sequence from the *Johann Friedrich Blumenbach: Bibliographie seiner Schriften* by Claudia Kroke. (See Appendix 1 for a TEI example).

TEI is an XML formatted document. For machines to read (i.e. "parse") XML, schema are used to globally identify the elements of the formatting. The schema definitions and rules for TEI are clearly documented and defined by the TEI-C. The TEI schema used with the 00027.xml file is formatted using the Relax NG language. Early in the project, I initiated the documentation of the TEI schema used in the working version of 000027.xml (see Appendix 2). I also used a tool known as "Roma" to create a new Relax NG schema document (see Appendix 3) that referenced the latest release (at that time) of the TEI standard known as TEI P5 v.2.5.0. The importance of validation of output in the NER annotation process is significant. Therefore, having a properly encoded XML schema is fundamental to this aspect of the project.

While the details of the TEI construction are beyond the scope of this paper, the basic XML element structure concept is worth noting. The basics are the *element* and the *attribute*.

For example: <pb n="**163**" facs="**#f0185**"/>

This is a page break element <pb> that marks the start of a new page. It has two attributes *n* and *facs* where *n* is the number (i.e. the page number) of the element and *facs* refers to the facsimile (the scanned page of text) that corresponds with the content of the element. These XML constructors consisting of element and attributes provide the building blocks for the structured text.

NER OBJECTIVES

1) Automated recognition of entities in the German texts of Blumenbach-online (so far: Handbook of Natural History in 12 editions).
2) Testing of NER strategies for historic texts (capacity building for the academy).

3) Generation of entities that allow a (more or less automated) linking via an ontology between texts and collection items and
4) Providing the results for the re-use in Blumenbach-Online

There are many complex problems inherent with attempting NER with historically specialized texts, but with a unique hybridized approach, the project has been able to target these objectives with specific methods:

- Usage of existing indexes and thematically similar word lists can facilitate the recognition and increase the recall.
- List enrichment via authority files (CERL, Getty, GND) allows for person identification.
- Specially adapted tools for correction and maintenance of the lists

A combination of list and rule-based NER seems most promising for this particular corpus of historical texts on natural history (Wettlaufer, 2013).

TIER 2: TRANSLATION

Referencing the architecture of the Semantic Web is essential to understand the translation tier. This is conceptually explained with this diagram.

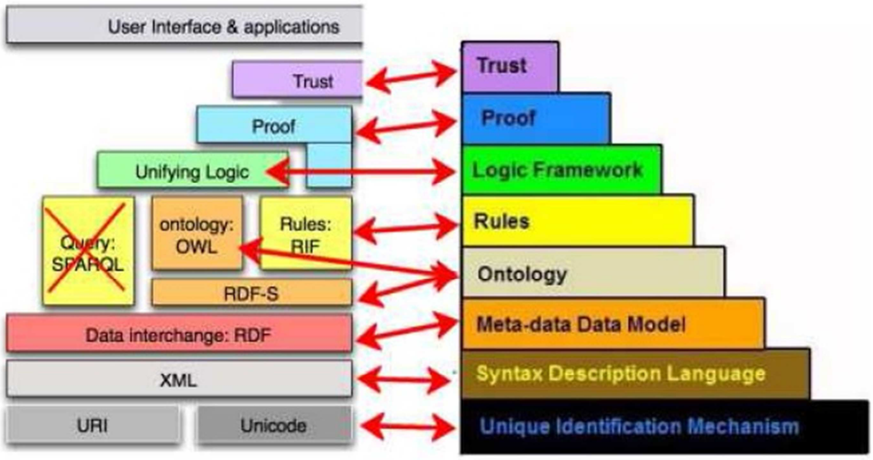


Figure 2: Semantic Web Model

Translation takes place above the XML-TEI (Syntax Description Language) and has three main layers: Meta-data Modelling, Ontology and Rules. The basis of the Meta-data Data Model is the use of RDF or Resource Description Framework.

RDF: RESOURCE DESCRIPTION FRAMEWORK

There are three types of objects in RDF: Resources, Properties and Statements. A resource is something that is described, a property is a specific aspect, characteristic, attribute, or relation used to describe resources and a statement is the expression that links the specific resource together with a named property plus the value of that property for that resource. These three individual parts of a statement are called, respectively, the subject, the predicate, and the object, and collectively referred to as an “RDF triple”.

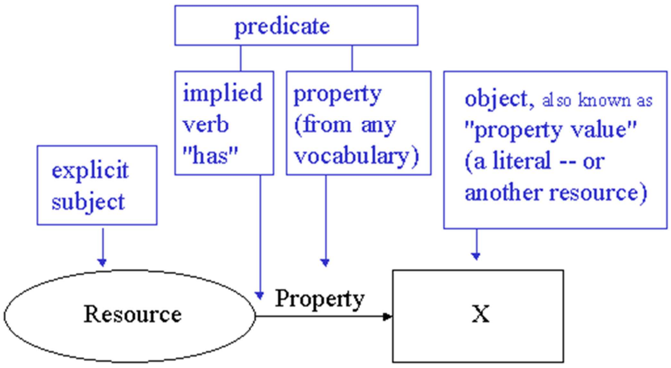

predicate
implied verb "has"
property (from any vocabulary)
object, also known as "property value" (a literal -- or another resource)
explicit subject
Resource
Property
X

Figure 3: An RDF Triple

For example, the expression "000027.xml documents *Handbuch der Naturgeschichte, 1799*" is information expressed in RDF syntax.

The principle objective of RDF is to represent named properties and property values with statements. How specific information, existing as a collection of triples, may be represented in practice is governed by an ontology, also called a conceptual reference model (CRM).

ONTOLOGY

What is ontology? In theory, ontology is "a formal, explicit specification of a shared conceptualisation" (Gruber, 1993). In practice, ontology is a logical model applied within a certain knowledge domain. The ontology chosen for Semantic Blumenbach is known as the CIDOC Conceptual Reference Model, designed particularly for concepts and information in cultural heritage and museum documentation.

"The CIDOC CRM is intended to promote a shared understanding of cultural heritage information by providing a common and extensible semantic framework that any cultural heritage information can be mapped to. It is intended to be a common language for domain experts and implementers to formulate requirements for information systems and to serve as a guide for good practice of conceptual modelling" (ICOM, 2014).

The language to describe the CIDOC ontology is known as OWL, or the Web Ontology Language. Similarly to the way that RNG describes XML schema, OWL describes an RDF schema that represents the ontology with a formal or canonical syntax. The Erlangen CRM OWL is an interpretation of the CIDOC CRM in a logical framework that follows the text of the specification. See Appendix 4 for a sample of the Erlangen OWL for CIDOC. Full Online Documentation of the Erlangen CRM OWL can be viewed at http://erlangen-crm.org/docs/ecrm/120111/index.html. The Erlangen CRM OWL provides the basis for the development of the Semantic Blumenbach localized OWL (see "ONTOLOGY MAPPING AND MODELING", page 10).

RULES

Represented significantly as the underlying component of the logic framework in the semantic web model in Figure 2, Rules define the interface between logic and the ontology. Why are Rules needed? There are applications where the ontology alone is not sufficient to logically express or represent the required knowledge domain. There are three fundamental components of knowledge representation in an OWL:

1. Axioms: the basic statements that an OWL ontology expresses.
2. Entities: elements used to refer to real-world objects.
3. Expressions: combinations of entities to form complex descriptions from basic ones. (W3, 2014)

Without the axiomatic or expressive aspects, an ontology may describe, but it cannot *reason*. The practical implementation of rule languages like SWRL (Semantic Web Rule Language) is an active and ongoing topic of many studies. The recently adopted recommendation for OWL 2 (http://www.w3.org/TR/owl2-overview/) reflects the significant contributions of "knowledge engineers" towards the advancement of semantic knowledge representation.

TIER 3: EXTENSION AND PRESENTATION

With specific elements and attributes from the XML-TEI properly translated to RDF triples, the presentation and extension of this data is the responsibility of an application framework. The extension component must be able to "ingest" the RDF triples, link them to a relational object repository and then extend the linked data for use in non-local presentation contexts. The presentation component should be a full-featured reader able to display the native historical text and the named entity annotations as hyperlinks. Additionally, the collections repository database should be closely integrated with the reader, facilitating the extension provided by the hyperlinked entities. A software framework known as WissKI (Wissenschaftliche Kommunikations Infrastruktur) has been identified as the prototype software platform for presentation and extension. (See WISSKI, p.12).

3. DOCUMENTATION AND EVALUATION

SEMANTIC WEB STUDY PROJECT WIKI

In order to organize and have a central point of reference for documentation, a secured area of my MediaWiki site <http://erbeinformatik.org/w/index.php?title=SP:Semantic_Blumenbach> was developed. As the documentation that was existing for the project was in several PDF files, and rather complicated and technical, the task of consolidating it into usable and editable content units was necessary. Additionally, the MediaWiki format enables direct hyperlinking of terms and resources for ready contextual referencing and keyword searching. As a research tool, MediaWiki is a very convenient way to organize different kinds of web based resources and also facilitates translation, in this case from German to English. For more information about MediaWiki, check this site <http://www.mediawiki.org>.

The first task at the beginning of the project was to understand how to analyze and read the Blumenbach structured text (000027.xml) file. I conducted a review of the current methods for rendering TEI, both natively and as readable text. Through this process, a general understanding of the state of the art of information extraction and named entity recognition software was developed.

NER METHODS EVALUATION

TEXT ENGINEERING OVERVIEW

The text engineering information extraction (NER) task of Semantic Blumenbach is one dimension of the broad scope of the field of Natural Language Processing (NLP). Semantic Blumenbach aims to provide semantic enrichment by linking collection objects to historical texts, but not to provide comparative linguistic analysis. Many text engineering functions can be applied to multiple text sources comparatively or in aggregate as a Corpora (a collection of similar texts). The tools for NER, therefore, need to be "right-sized" and customized for a particular objective. The Semantic Blumenbach team has, in fact, developed custom software to their task that combines existing "rule-based" and "list-based" NER tools (Wettlaufer, 2013).

Specialized software tools are essential in both the original creation and subsequent annotation of a TEI file with a Named Entity Recognition process. It is important to note that with a historical text corpus, prerequisite information extraction functions provide the foundation for an accurate Named Entity Recognition process. In a list-based tool, a database of extracted keywords provides a dictionary for the NER software and these keywords must match the unique attributes of the text to be annotated, including spelling and language. For an 18th century scientific text in German, developing this dictionary is complex task in its own domain.

As mentioned previously and illustrated in Appendix 1, TEI is an XML format. Though considered "human readable", a structured text document obscures the text source content with markup tags (e.g. elements and attributes). A tool known as a "reader" is required to render XML (i.e. to filter the XML tags and present the text source with legible formatting). However, simply reading the text is the most basic function of the information extraction process of text engineering. Other common text engineering information extraction tasks include tokenizing, sentence splitting, part of speech tagging (lemmatization), concordance, co-occurrence, progression analysis and referencing.

Therefore, investigation of the principle tools of text engineering was important for developing a more complete view of the Semantic Blumenbach project. While not a complete survey of what tools were evaluated, the four tools listed in the following sections comprise the substantive and core functionality needed for basic NER.

GATE DEVELOPER

GATE Developer is an integrated development environment that provides a set of graphical interactive tools for the creation, measurement and maintenance of software components for processing human language. GATE, or "General Architecture for Text Engineering", is a set of tools that assist with functionality relating to the following core concepts:

- the documents to be annotated
- Corpora comprising sets of documents, grouping documents for the purpose of running uniform processes across them
- annotations that are created on documents
- annotation types such as 'Name' or 'Date'
- annotation sets comprising groups of annotations
- processing resources that manipulate and create annotations on documents
- applications, comprising sequences of processing resources, that can be applied to a document or corpus.

The specific applications of GATE are diverse and complex. The core logic of the application can be characterized as using "machine learning" algorithms for the construction of rule-based classification systems. The Semantic Blumenbach researchers determined that GATE's method of rule based NER processing proved difficult when it was applied to historical German texts (Wettlaufer, 2013).

TEXTGRID

TextGrid <http://www.textgrid.de/en/> is a "Virtual Research Environment for the Humanities". It primarily provides a repository infrastructure for digital editions. The editions are managed with similar methods known in software development. In fact, the TextGrid Laboratory interface uses the Eclipse Integrated Development Environment. Conceptually, the repository paradigm is well-designed, but practically the authentication and collaboration mechanisms are somewhat problematic. TextGrid assumes that researchers will be linked into the D-Grid (DFN), but BTU has not subscribed to this network. I was however able by means of my SLUB (Staats- und Universitätsbibliothek Dresden) ID card to gain access. Once you have authenticated, many resources are available, including thousands of pages of German language historical texts, many in TEI format.

OXYGEN

According to its website “<oXygen/> is the best XML editor available”. It is a robust commercial software platform that could be considered the market leader for Visual XML Editing. I evaluated the software with the 30-day trial license and found that it had many valuable features, including but not limited to, schema documentation, schema validation, TEI Editing, Relax NG editing, and XSLT Editing (described in TRIPLIFYING WITH XSLT below).

The 000027.xml schema referenced in Appendix 2 is a sample of the Oxygen schema documenting function. The entire document is published here http://erbeinformatik.org/sp/sbschema.html.

The most useful function of Oxygen for the purpose of contributing to the Semantic Blumenbach project was the schema validation. This checks the structure of the XML against the schema source, in our case, a Relax NG file. I initiated a two-phase validation of the Blumenbach TEI schema. First, I created the most basic TEI schema with the Roma tool (http://www.tei-c.org/Roma/) and in successive passes of validation with Oxygen, added the TEI modules that comprise the 000027.xml file and its NER annotations. Several problems with the 000027.xml were identified and noted to GCDH as follows:

1) the <ornament> element is invalid. This should be replaced with <figure> and in a few cases <figDesc>
2) the sortKey attribute values will not work with a "separator space" because of a regular expression validation ["(\p{L}|\p{N}|\p{P}|\p{S})+"] that does not allow them.
3) the <date to="date" from="date"> does not work in 2.5. [<date min="date" max="date"> is now favored instead.]

These problems resulted from the de facto TEI schema for the TextGrid version of the 000027.xml that referenced the 1.9 release. The end result of this process was an updated Relax NG file that includes the 2.5 release of the TEI specification. Also, a Roma source file (known as an ODD) for TEI schema alterations was provided to GCDH.

TEXTOMETRIE

Textometrie is a discipline developed primarily in France since the 1970s and involves statistical lexical evaluation of the rich vocabulary of a text. It continues the data analysis (factor analysis, classifications) methods developed by Jean-Paul Benzécri (1973) applied to linguistic data. These techniques are used to generate synthetic and visual mapping of words and texts as they are related or opposed in a corpus. Calculation results are synthetic, selective and suggestive reorganizations. The interpretation of the calculations are based on quantitative indicators but also on the systematic review of contexts, now facilitated by relevant hyperlinks.

A new application for textometrie called TXM provided a very interesting platform for the linguistic analysis of the Blumenbach text as well as enabling information extraction. With it, I generated a lexicon of the entire 000027.xml file. The lexicon includes a word frequency metric, so the number of times a particular word was used can be measured. The most common word is “und” used 2594 times and the most common entity is “Erde” used 342 times. Additionally, I generated a lemmatized (part of speech tagged) list of the entire lexicon. The partitioning or filtering of the text into parts of speech is a basic function of NER, as “entities” are always nouns.

Importantly, TXM is a robust tool that includes a scripting engine. The basis of this engine is called Groovy, which is an object-oriented language for the Java platform. Groovy, like Java, can do most of the processing “work” of NER that includes data extraction and validation, described in the next section.

DATA EXTRACTION AND VALIDATION

The final process with NER is validation where the TEI annotations sourced from the NER software process are extracted and validated for accuracy and completeness. Based on the understanding that XML-TEI file structured texts could be "mined" for data, several different XML database query (XQUERY) methods for accomplishing this were reviewed.

The power of Groovy (and Java) for NER tasks with XML-TEI cannot be understated. In order to understand how NER works for data extraction, I experimented with writing a script that could extract values from specific TEI elements on specific XML node paths. This involved familiarity with the principle query language for XML known as XPATH. See Appendix 6 for an example of the output of this script experiment.

A very nice tool called BaseX proved to be the most useful extractor. BaseX uses a tabular representation of XML tree structures to store XML documents. The database acts as a container for a single document or a collection of documents. The XPath Accelerator encoding scheme and Staircase Join Operator have been taken as inspiration for speeding up XPath location steps (Grün, 2011).

TRANSLATION METHODS

The purpose of creating and developing the structured text file with NER is to facilitate the output of linked open data with an automated process. The foundation of LOD is RDF. The creation of RDF from XML input is done with successive methods. The primary method is mapping where the XML elements are mapped to resource classes defined in the ontology. The secondary method is "triplifying" which is creating the statements that exist as RDF triples by connecting resources together with a property and outputting them in a serialized format.

ONTOLOGY MAPPING AND MODELING

A significant task in translation of specific XML.-TEI elements to RDF triples is the creation of paths. A path is an order of relationships that describes dependency, inheritance and hierarchy. With CIDOC, the class and property relationships are very elaborate. An adaptation of the Erlangen CRM OWL for the Semantic Blumenbach project is being developed that attempts to explain the logic that connects the NER related TEI element and attributes as RDF triples. My research contribution for this project component was to explore the modeling potential of different path relationships. This was accomplished by referencing other CIDOC mapping examples and developing graphs with the Protégé OWL Editor.

PROTÉGÉ-OWL EDITOR

The Protégé-OWL editor enables users to:

- Load and save OWL and RDF ontologies.
- Edit and visualize classes, properties, and SWRL rules.
- Define logical class characteristics as OWL expressions.
- Execute reasoners such as description logic classifiers.
- Edit OWL individuals for Semantic Web markup.

See Appendix 5 for a graph of the CIDOC document class (E31) that I developed with the Protégé OWL Editor. This graph was based on the 2011 work of Bountouri & Gergatsoulis that attempts to map the EAD (Encoded Archival Description) model to CIDOC. As previously mentioned, the tasks of mapping, modeling and graphing ontology is considered "knowledge engineering".

Noy and McGuinness (2001) defined a method for knowledge engineering with the following steps:

1. Determine the domain and scope of the ontology
2. Consider reusing existing ontologies
3. Enumerate important terms in the ontology
4. Define the classes and the class hierarchy
5. Define the properties of classes—slots
6. Define the facets of the slots
7. Create instances

The scope of research for my work was primarily concerned with class hierarchy. A class hierarchy can be defined with a "top-down" or a "bottom-up" approach. In the case of Semantic Blumenbach, a combined approach was required because CIDOC is a "top-down" ontology, but NER is a "bottom-up" technique. The main issue that was realized in the development of bottom-up path logic to connect TEI elements together with CIDOC was the need for axioms. This is due to the fact that CIDOC allows multiple inheritance, does not restrict class inheritance with transitive properties, and class relationships often span multiple steps. Here is an example encountered when modelling the TEI <term> element from a top-down approach:

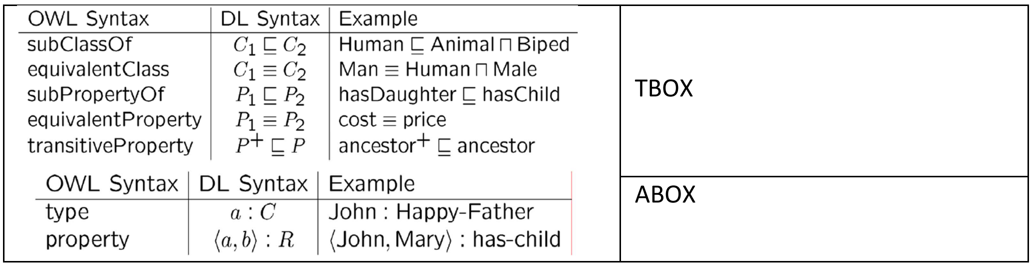
- <term>Fische</term>: E31 Document → P106 is composed of → E31 Document → P70 documents → E22 Man-Made Object → P128 carries → E73 Information Object [000027.xml] → P67 refers to → E75 Conceptual Object Appellation [<term>]→ P67 refers to →E33 Linguistic Object [Fische]
  and
- E73 Information Object [000027.xml] → P67 refers to → E75 Conceptual Object Appellation [<term>] → P2 has type --> E55 Type[TEI 2.5.0.model.emphLike]--> P71i is listed in --> E32 Authority Document [http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-term.html]

The TEI element <term> has at least two branches from the top-down, because it is in the domain of both the Linguistic Object property "P67 refers to" and the parent domain of the Authority Document property "P71i is listed in". Thus, defining a bottom-up or inverse path for <term>Fische</term> that similarly connects these CIDOC classes is not possible. Description Logic with an axiomatic inference system must therefore be provided in order to fully express XML-TEI elements as RDF resources in the OWL implementation of the CIDOC CRM. The application of DL in an ontology is known as reasoning.

REASONING

OWL relies on what is known as "the open world assumption". This basically means that by the absence of a statement, a reasoner cannot (and must not) infer that the statement is false. Description Logic reasoners use what is known as the method of analytic tableaux. In proof theory, the semantic tableau (or truth tree) is a decision procedure for sentential and related logics, and a proof procedure for formulas of first-order logic. Tableau calculus consists of a collection of rules with each rule specifying how to break down one logical connective into its constituent parts (Letz and Stenz, 2001).

In DL, a distinction is drawn between the so-called TBox (terminological box) and the ABox (assertional box). In general, the TBox contains sentences describing concept hierarchies (i.e., relations between concepts) while the ABox contains ground sentences stating where in the hierarchy individuals belong (i.e., relations between individuals and concepts).



OWL Syntax	DL Syntax	Example
subClassOf	$C_1 \sqsubseteq C_2$	Human ⊑ Animal ⊓ Biped
equivalentClass	$C_1 \equiv C_2$	Man ≡ Human ⊓ Male
subPropertyOf	$P_1 \sqsubseteq P_2$	hasDaughter ⊑ hasChild
equivalentProperty	$P_1 \equiv P_2$	cost ≡ price
transitiveProperty	$P^+ \sqsubseteq P$	ancestor$^+$ ⊑ ancestor

TBOX

OWL Syntax	DL Syntax	Example
type	$a : C$	John : Happy-Father
property	$\langle a, b \rangle : R$	⟨John, Mary⟩ : has-child

ABOX

Figure 4 TBox and Abox Axioms

In logic, a rule of inference, inference rule, or transformation rule is a logical form consisting of a function which takes premises, analyzes their syntax, and returns a conclusion (or conclusions). The most common decision problems are basic database-query-like questions (instance checking, relation checking, subsumption and concept consistency).

The key discussion about DL reasoning in ontologies revolves around decidability. In practice, a conclusion cannot be computed in an undecidable ontology. The objective with DL is to create decidability with rules that can be confirmed with reasoning.

A complete discussion of DL reasoning is beyond the scope of this paper. It is important to note that this aspect of the semantic web is perhaps the most prominent obstacle in its broader implementation. In relation to the OWL model for Semantic Blumenbach, DL rules should be carefully considered.

TRIPLIFYING WITH XSLT

Even without a fully DL expressive OWL for the Semantic Blumenbach implementation of CIDOC, RDF triples can still be created (serialized and formatted) from an XML-TEI annotated file. As mentioned, this is a secondary method that follows the mapping of classes. One mechanism for "triplifying" is with an XSLT or Extensible Stylesheet Language Transformation. The XSLT is applied on an XML source and processed using a tool like Oxygen to generate an output file. The transformation is a normal operation that is not that unlike the "find and replace" (regular expression) function of most word processors. The source XML-TEI file is parsed for occurrences of a specific element like <term>, and the attributes of the element are replaced with classes or properties. The replaced attributes are then concatenated into an object instance (the triple).

Martin Scholz's XSLT triplifying script can be seen here:
<https://github.com/mnscholz/wisski_texttei/blob/master/triplify.xsl>

In the case of Semantic Blumenbach, the serialized output file consisting of RDF triples is then capable of being "ingested" into the WissKI system.

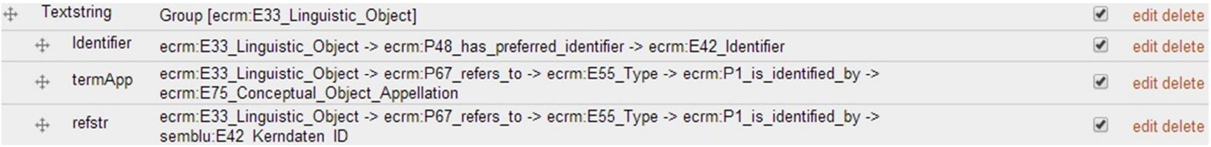
EXTENSION AND PRESENTATION FRAMEWORK

WISSKI

WissKI is described as a "communications platform for curated knowledge." WissKI implements an RDF storage facility, called a "Triple Store", built entirely on semantic web technology that enables the creation of new methods for scientific workflow and content management. It is also the foundation of a semantic text annotation mechanism that allows the "ingest" of structured information to the system. The text is analysed to connect mentioned entities (names, places, dates etc.) to the systems knowledge base. The concept of WissKI, therefore, is perceived by Semantic Blumenbach as having substantial value towards the project objectives.

WissKI currently exists as several extension modules written for the Drupal 6 content management platform. The main RDF storage functionality of the WissKI core is built on ARC2, a flexible RDF system written in PHP. ARC2 also includes a SPARQL endpoint class that allows for the remote retrieval and manipulation of the ingested RDF data by means of queries. The SPARQL endpoint is perceived as the main export mechanism for the production of linked open data.

An interface to the triple store is provided through forms. The core forms field constructor of the WissKI system is provided with the "pathbuilder" module. This tool allows the administrator to construct semantic definitions for the content creation of the system based on the loaded ontology.



Textstring	Group [ecrm:E33_Linguistic_Object]	☑	edit delete
Identifier	ecrm:E33_Linguistic_Object -> ecrm:P48_has_preferred_identifier -> ecrm:E42_Identifier	☑	edit delete
termApp	ecrm:E33_Linguistic_Object -> ecrm:P67_refers_to -> ecrm:E55_Type -> ecrm:P1_is_identified_by -> ecrm:E75_Conceptual_Object_Appellation	☑	edit delete
refstr	ecrm:E33_Linguistic_Object -> ecrm:P67_refers_to -> ecrm:E55_Type -> ecrm:P1_is_identified_by -> semblu:E42_Kerndaten_ID	☑	edit delete

Figure 5 WissKI Pathbuilder Screenshot

The WissKI pathbuilder definitions are used to map the fields for data aggregation in the system.   Martin Scholz, a lead developer of WissKI, describes this method as follows “for each field of a form you have to define a path in the pathbuilder.

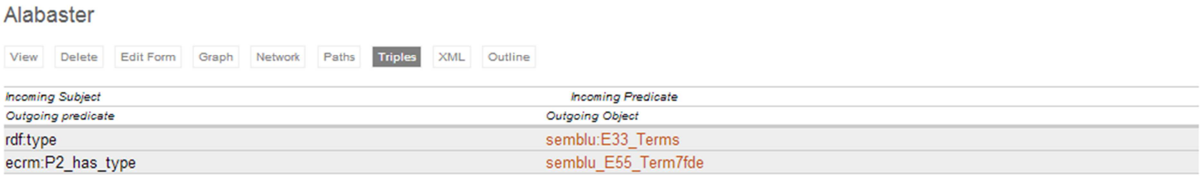
Each path has the following structure:

Class0 -> ObjectProperty0 -> Class1 -> ObjectProperty1 -> ... -> ClassN -> DatatypeProperty

Triples are created when filling a field with value <value> and saving the form. For each step ClassX -> ObjectPropertyX -> ClassY, 3 triples with 2 instances are created (if they do not exist already):

- instX rdf:type ClassX
- instY rdf:type ClassY
- instX ObjectPropertyX instY

For the last step, only the following triple is created: instN DatatypeProperty <value>”   Below is a <term> example triple with the value <Alabaster> created from the termApp path in the above screenshot.

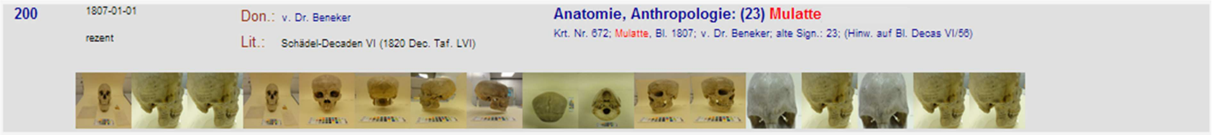


Incoming Subject	Incoming Predicate
Outgoing predicate	*Outgoing Object*
rdf:type	semblu:E33_Terms
ecrm:P2_has_type	semblu_E55_Term7fde

Figure 6 Term Alabaster as a WissKI Triple

Note that the WissKI system appends unique identifiers to the class label to create the specific instance object.

PROOF OF CONCEPT EXAMPLE



200

1807-01-01

rezent

Don.: v. Dr. Beneker

Lit.: Schädel-Decaden VI (1820 Dec. Taf. LVI)

Anatomie, Anthropologie: (23) Mulatte

Krt. Nr. 672; Mulatte, Bl. 1807; v. Dr. Beneker; alte Sign.: 23; (Hinw. auf Bl. Decas VI/56)

Figure 7 Mulatte Resource from Blumenbach Online Repository

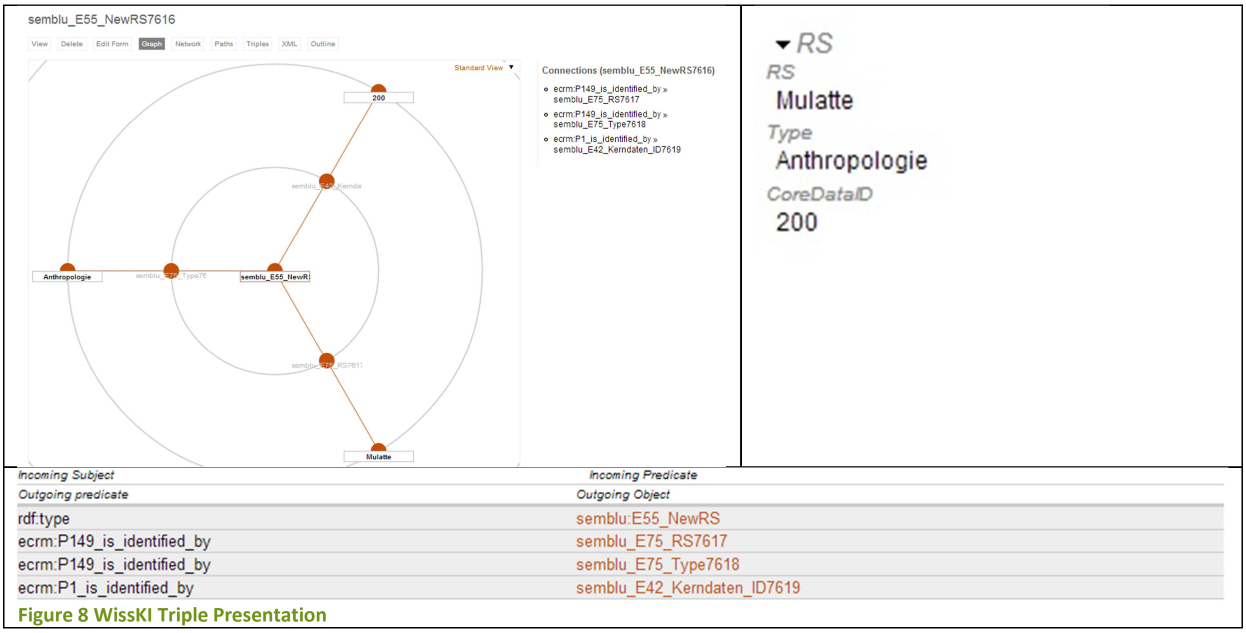
Starting with the metadata from the skull object described in the Blumenbach Online collection, the primary objective of describing the Blumenbach Online collection objects within the TEI text has been achieved through the use of the TEI element <rs> with the attribute “ref” that is the collection repository identifier.

<rs type="Anthropologie" ref="**200**">Mulatte</rs> (From 000027.xml, line 8873 [Register])

The metadata ID “200” is annotated with NER in the TEI element <rs> as a ref attribute in the example above.

In the WissKI system, the presentation environment aims to hyperlink the natively displayed Blumenbach text with Blumenbach Online data and images catalogued in a standard relational database by retrieving the RS ref ID from the Triple store.

The triple for this collection object is ingested into the system and can be viewed in WissKI as follows:



Connections (semblu_E55_NewRS7616)

- ecrm:P149_is_identified_by » semblu_E75_RS7617
- ecrm:P149_is_identified_by » semblu_E75_Type7618
- ecrm:P1_is_identified_by » semblu_E42_Kerndaten_ID7619

▾ RS

RS
Mulatte

Type
Anthropologie

CoreDataID
200

Incoming Subject	*Incoming Predicate*
Outgoing predicate	*Outgoing Object*
rdf:type	semblu:E55_NewRS
ecrm:P149_is_identified_by	semblu_E75_RS7617
ecrm:P149_is_identified_by	semblu_E75_Type7618
ecrm:P1_is_identified_by	semblu_E42_Kerndaten_ID7619

Figure 8 WissKI Triple Presentation

The presentation environment (also known as the “front end”) remains a work in progress. Currently, WissKI can display ingested TEI through Drupal’s “book format”, but the reader functions are limited.

The find functionality of WissKI provides an ability to search for specific entities (person, place, term, etc. Currently, WissKI uses a module that adds an OPAC (Online Public Access Catalogue) type search facility for searching the triple store. Search masks are generated using the paths defined via the pathbuilder module.

4. CURRENT PROJECT STATUS

With the completion of the NER tier at the beginning of January 2014, the project focus is now on translating the XML-TEI to RDF. The status quo process for this depends primarily on a clear ontology OWL schema for the WissKI system. The finalization of the Semantic Blumenbach OWL is a work in progress. There does not appear to be a generic “converter” for this process, and it remains a highly localized task. The current XSLT triplifier requires specific class mapping within the script and does not accept parameters or a configuration file. Optimally, development of a more generic approach for translation will remain a project objective.

The current version of WissKI was built while the project was funded from 2009-2011. Since 2011, development has been limited and thus several important deprecation issues have not been addressed. One of the main issues is that Drupal 6 is not compliant with new PHP standards. WissKI has, however, received a new round of funding recently. While it appears that the development of WissKI is a constraint to the progression of Semantic Blumenbach, the important conceptual groundwork for future practical implementation has clearly been established. WissKI is a great concept and if implemented with long-term versioning objectives in an upgradeable framework should continue to remain viable for the near future.

As revealed through this research, the technical challenges of knowledge representation and knowledge engineering in the semantic web are substantial. Reflecting on the Semantic Blumenbach project objectives, providing a proof of concept (objective 1) is readily achievable. Objective 2, “publication of Linked Open Data”, is currently possible with SPARQL, yet the extensibility and value of triples that are constrained by locally defined unique identifiers is questionable. And finally, objective 3, developing generic tools for projects in the academy of science, remains a programming task where reusable libraries are created for different semantic tasks, like RDF triplifiication and ontology mapping and path building.

5. CONCLUSION

The theoretical premise of this independent study project was that the use of Kulturinformatik, as represented in the semantic technologies of Semantic Blumenbach, could be applied for intelligent transdisciplinary heritage information retrieval and analysis. While this possibility has not been demonstrated, it remains a viable extension of the research presented here. Through the use of the tools and methods introduced in Semantic Blumenbach, new possibilities for heritage research can be considered.

The objectives of the study project as outlined in the initial proposal have therefore been achieved. These were as follows:

1. Document project framework, standards, reference models, and code base.
2. Evaluate application tools, techniques and methods.
3. Perform problem-solving and development tasks as directed by Dr. Jörg Wettlaufer, Researcher in the Digital Humanities Research Collaboration (DHFV).

For a complete correspondence history of the study project, please review Appendix 7.

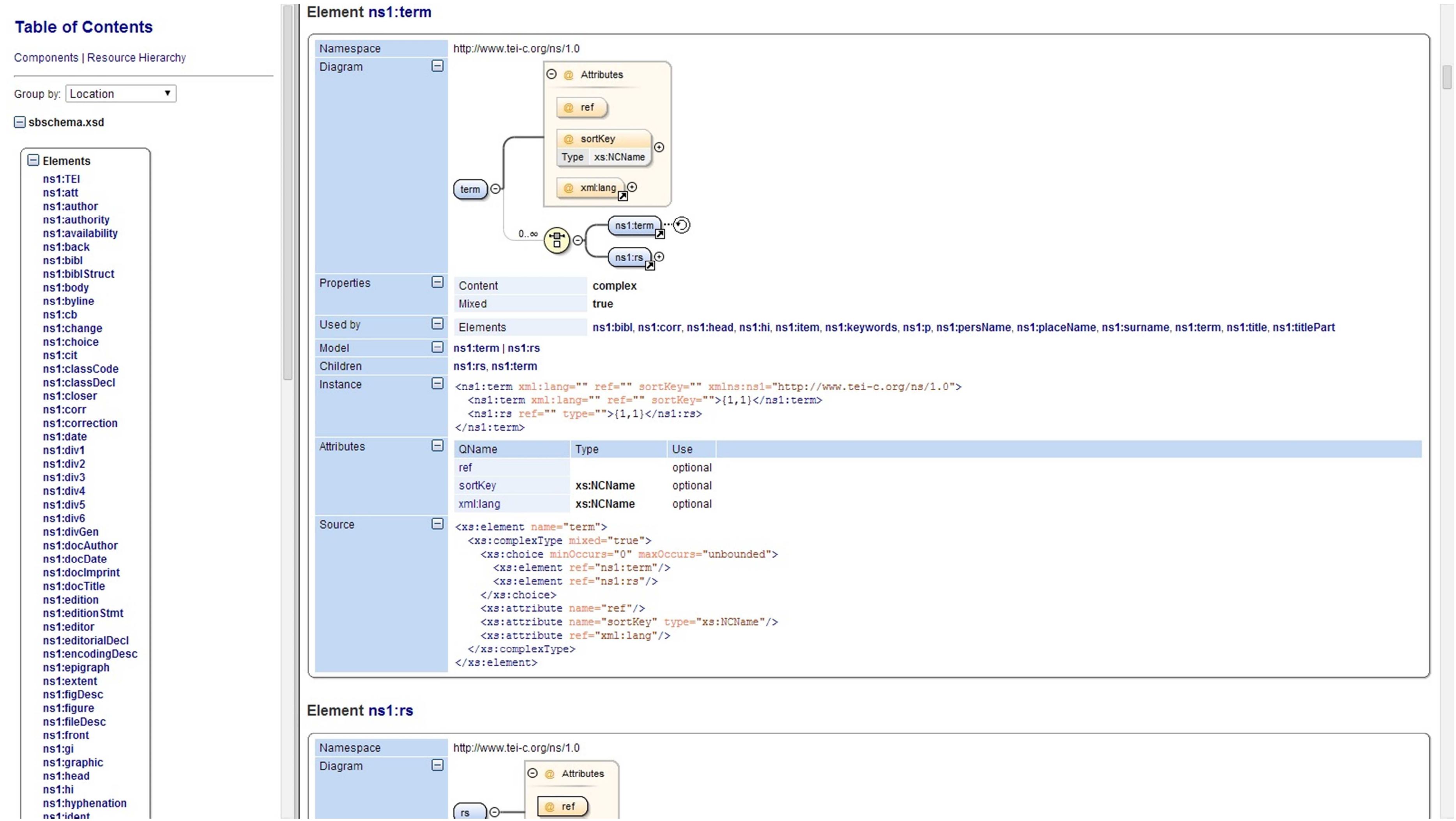
APPENDICES

APPENDIX 1: TEI EXAMPLE

Excerpt from 000027.xml

```
<p rendition="#l1em">
        In <placeName ref="#GettyId:7005685">Canada</placeName>
        , auf <placeName ref="#GettyId:7013071">Labrador</placeName>
        , um die<placeName ref="#GettyId:7013052"> Hudsonsbay </placeName>
        etc. Thut zumahl im Winter den jungen<lb/>
        Baumstmmen groen Schaden.
</p>
<p rendition="#indent-2">
        2. <hi rendition="#i">
                <hi rendition="#r">
                        <term xml:lang="la" SortKey="Hystrix_Cristata">Cristata</term>
                </hi>
        </hi>.
        <hi rendition="#r">H. spinis longissimis, capite cristato, cauda abbreuiata</hi>
.</p>
<p rendition="#l2em">
v.
        <persName xml:lang="de" ref="http://thesaurus.cerl.org/record/cnp01362609">
                <surname>Schreber</surname>
        </persName>
        <hi rendition="#r">tab</hi>
        . 167
.</p>
<p rendition="#l1em">
Ursprnglich im wrmern
<placeName ref="#GettyId:1000004">Asien</placeName>
 und fast ganz
<lb/>
<placeName ref="#GettyId:7001242">Africa</placeName>
; nhrt sich zumahl von Baumrinden; nistet
<lb/>
in die Erde. Im Zorn rasselt es mit seinen Stacheln, die ihm zuweilen, besonders im Herbst,
<lb/>ausfallen; kann sie aber nicht gegen seine Verfolger von sich schieen!
<note n="*)" anchored="true" place="bottom">
        <pb xml:id="pb084_0002" n="84" facs="images/00000108"/>
        <p>
                Der weiland als Panazee berufne thierische Gallenstein (
                <hi rendition="#i">
                        <hi rendition="#r">piedra del porci</hi>
                </hi>
                ) soll sich in einer noch
                <lb/>
                nicht genau bekannten ostindischen Gattung von
                <lb/>
                Stachelschweinen finden.
        </p>
</note></p>
```

APPENDIX 2: TEI SCHEMA DOCUMENTATION FROM OXYGEN TOOL



Element ns1:term

Namespace: http://www.tei-c.org/ns/1.0

Diagram ⊟

Properties ⊟

Content	complex
Mixed	true

Used by ⊟

Elements	ns1:bibl, ns1:corr, ns1:head, ns1:hi, ns1:item, ns1:keywords, ns1:p, ns1:persName, ns1:placeName, ns1:surname, ns1:term, ns1:title, ns1:titlePart

Model ⊟ ns1:term | ns1:rs

Children ns1:rs, ns1:term

Instance ⊟

```
<ns1:term xml:lang="" ref="" sortKey="" xmlns:ns1="http://www.tei-c.org/ns/1.0">
  <ns1:term xml:lang="" ref="" sortKey="">{1,1}</ns1:term>
  <ns1:rs ref="" type="">{1,1}</ns1:rs>
</ns1:term>
```

Attributes ⊟

QName	Type	Use
ref		optional
sortKey	xs:NCName	optional
xml:lang	xs:NCName	optional

Source ⊟

```
<xs:element name="term">
  <xs:complexType mixed="true">
    <xs:choice minOccurs="0" maxOccurs="unbounded">
      <xs:element ref="ns1:term"/>
      <xs:element ref="ns1:rs"/>
    </xs:choice>
    <xs:attribute name="ref"/>
    <xs:attribute name="sortKey" type="xs:NCName"/>
    <xs:attribute ref="xml:lang"/>
  </xs:complexType>
</xs:element>
```

Element ns1:rs

Namespace: http://www.tei-c.org/ns/1.0

Diagram ⊟

Complete Documentation available online: at <http://erbeinformatik.org/sp/sbschema.html>

Appendix A.1.199 <term>

<term> contains a single-word, multi-word, or symbolic designation which is regarded as a technical term. [3.3.4.]	
Module	core
Attributes	Attributes att.global (@xml:id, @n, @xml:lang, @rend, @rendition, @xml:space) (att.global.linking (@corresp, @next, @prev)) (att.global.analytic (@ana)) (att.global.facs (@facs)) att.declaring (@decls) att.pointing (@targetLang, @target, @evaluate) att.typed (@type, @subtype) att.canonical (@key, @ref) att.sortable (@sortKey) att.cReferencing (@cRef)
Member of	model.emphLike
Contained by	**analysis**: s **core**: abbr add addrLine author bibl biblScope corr date del desc editor emph expan foreign gloss head hi index item l label mentioned name note num orig p pubPlace publisher q quote ref reg resp rs sic soCalled speaker stage term time title unclear **figures:** cell figDesc **header**: authority catDesc change classCode creation distributor edition extent funder keywords language licence principal rendition sponsor tagUsage **linking:** seg **namesdates:** addName affiliation age birth bloc country death district education faith floruit forename genName geogFeat geogName langKnown nameLink nationality occupation orgName persName placeName region residence roleName settlement sex socecStatus surname **tagdocs:** eg **textstructure:** byline closer dateline docAuthor docDate docEdition docImprint imprimatur opener salute signed titlePart trailer
May contain	**analysis**: interp interpGrp pc s w **core:** abbr add address cb choice corr date del emph expan foreign gap gloss graphic hi index lb mentioned milestone name note num orig pb ptr ref reg rs sic soCalled term time title unclear **figures:** figure formula **header:** idno

	linking: anchor seg **namesdates:** addName affiliation bloc climate country district forename genName geo geogFeat geogName location nameLink offset orgName persName placeName population region roleName settlement state surname terrain trait **tagdocs:** att code gi ident tag val
Declaration	element term { att.global.attributes, att.declaring.attributes, att.pointing.attributes, att.typed.attributes, att.canonical.attributes, att.sortable.attributes, att.cReferencing.attributes, macro.phraseSeq }
Example	A computational device that infers structure from grammatical strings of words is known as a **<term>**parser**</term>**, and much of the history of NLP over the last 20 years has been occupie with the design of parsers.
Example	We may define **<term** **rend**="sc" **xml:id**="TDPV"**>**discoursal point of view**</term>** as **<gloss** **target**="#TDPV"**>**the relationship, expressed through discourse structure, between the implied author or some other addresser, and the fiction.**</gloss>**
Note	This element is used to supply the form under which an index entry is to be made for the location of a parent <index> element.In formal terminological work, there is frequently discussion over whether terms must be atomic or may include multi-word lexical items, symbolic designations, or phraseological units. The <term> element may be used to mark any of these. No position is taken on the philosophical issue of what a term can be; the looser definition simply allows the <term> element to be used by practitioners of any persuasion. As with other members of the **att.canonical** class, instances of this element occuring in a text may be associated with a canonical definition, either by means of a URI (using the ref attribute), or by means of some system-specific code value (using the key attribute). Because the mutually exclusive target and cRef attributes overlap with the function of the ref attribute, they are deprecated and may be removed at a subsequent release.

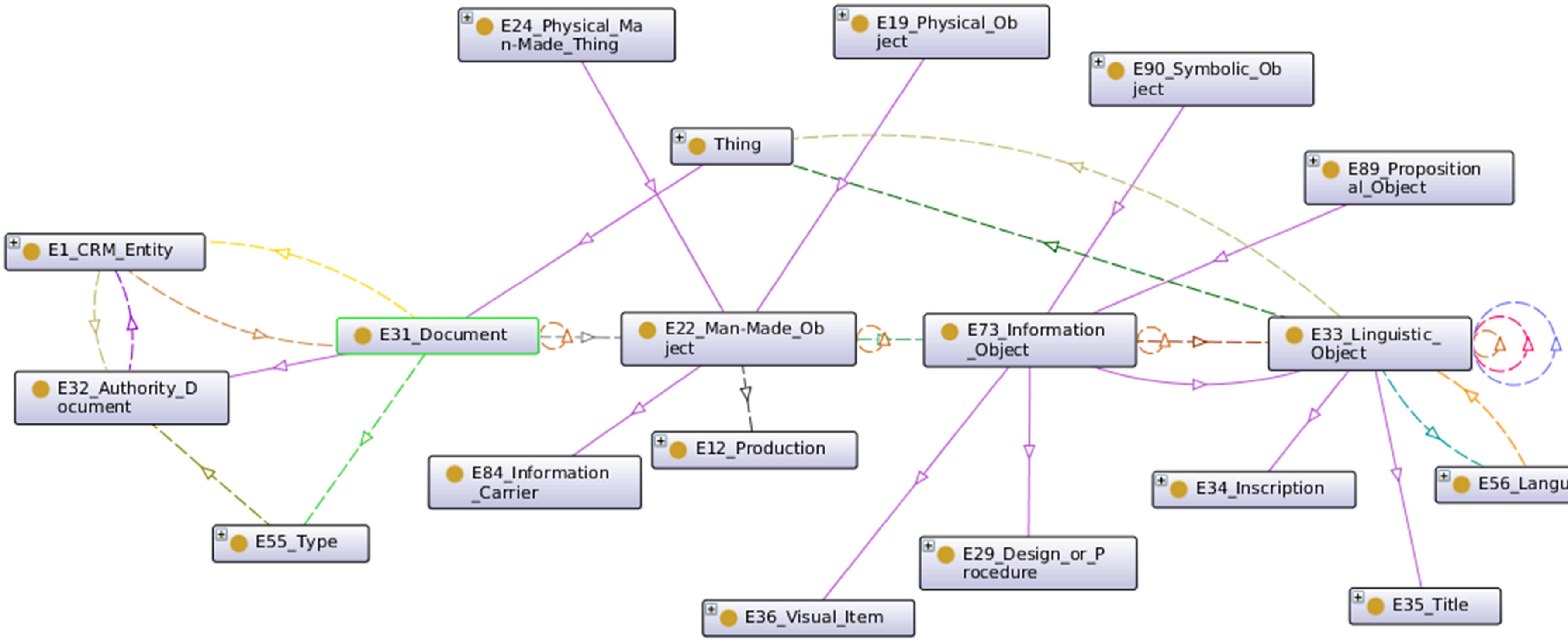
Complete Roma generated documentation available online: <http://erbeinformatik.org/sp/blumenbach_tei_lite_doc.html>

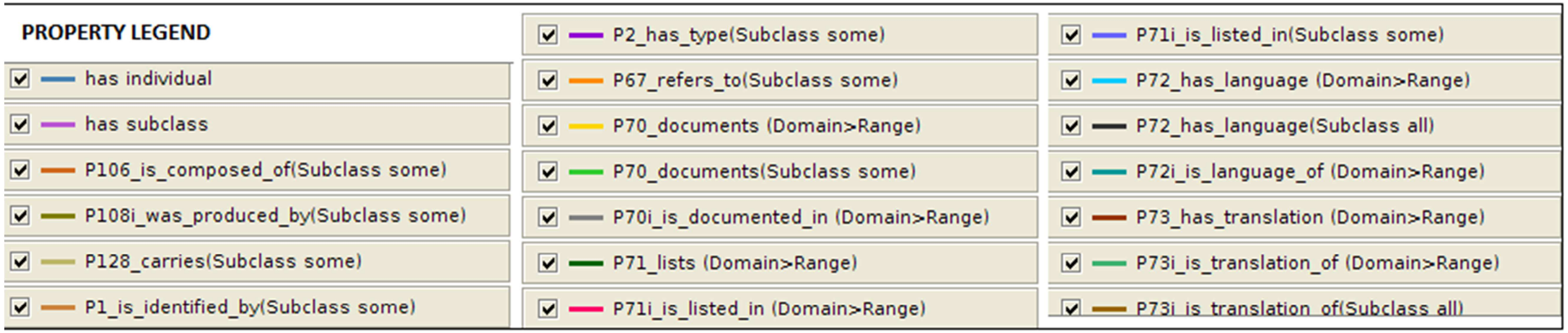
APPENDIX 4: ERLANGEN OWL

This is an excerpt from the OWL file that describes the CIDOC CRM with RDF. Specifically, noted here are the properties P70 through P73. Each property is also listed with an inverse syntax noted with the *i* (e.g. P70i).

```
<ObjectPropertyDomain>
    <ObjectProperty abbreviatedIRI="ecrm:P70_documents"/>
    <Class abbreviatedIRI="ecrm:E31_Document"/>
  </ObjectPropertyDomain>
  <ObjectPropertyDomain>
    <ObjectProperty abbreviatedIRI="ecrm:P70i_is_documented_in"/>
    <Class abbreviatedIRI="ecrm:E1_CRM_Entity"/>
  </ObjectPropertyDomain>
  <ObjectPropertyDomain>
    <ObjectProperty abbreviatedIRI="ecrm:P71_lists"/>
    <Class abbreviatedIRI="ecrm:E32_Authority_Document"/>
  </ObjectPropertyDomain>
  <ObjectPropertyDomain>
    <ObjectProperty abbreviatedIRI="ecrm:P71i_is_listed_in"/>
    <Class abbreviatedIRI="ecrm:E1_CRM_Entity"/>
  </ObjectPropertyDomain>
  <ObjectPropertyDomain>
    <ObjectProperty abbreviatedIRI="ecrm:P72_has_language"/>
    <Class abbreviatedIRI="ecrm:E33_Linguistic_Object"/>
  </ObjectPropertyDomain>
  <ObjectPropertyDomain>
    <ObjectProperty abbreviatedIRI="ecrm:P72i_is_language_of"/>
    <Class abbreviatedIRI="ecrm:E56_Language"/>
  </ObjectPropertyDomain>
  <ObjectPropertyDomain>
    <ObjectProperty abbreviatedIRI="ecrm:P73_has_translation"/>
    <Class abbreviatedIRI="ecrm:E33_Linguistic_Object"/>
  </ObjectPropertyDomain>
  <ObjectPropertyDomain>
    <ObjectProperty abbreviatedIRI="ecrm:P73i_is_translation_of"/>
    <Class abbreviatedIRI="ecrm:E33_Linguistic_Object"/>
</ObjectPropertyDomain>
```

APPENDIX 5: CLASS ORDER RELATIONSHIP GRAPH FROM CIDOC




PROPERTY LEGEND

- ☑ has individual
- ☑ has subclass
- ☑ P106_is_composed_of(Subclass some)
- ☑ P108i_was_produced_by(Subclass some)
- ☑ P128_carries(Subclass some)
- ☑ P1_is_identified_by(Subclass some)
- ☑ P2_has_type(Subclass some)
- ☑ P67_refers_to(Subclass some)
- ☑ P70_documents (Domain>Range)
- ☑ P70_documents(Subclass some)
- ☑ P70i_is_documented_in (Domain>Range)
- ☑ P71_lists (Domain>Range)
- ☑ P71i_is_listed_in (Domain>Range)
- ☑ P71i_is_listed_in(Subclass some)
- ☑ P72_has_language (Domain>Range)
- ☑ P72_has_language(Subclass all)
- ☑ P72i_is_language_of (Domain>Range)
- ☑ P73_has_translation (Domain>Range)
- ☑ P73i_is_translation_of (Domain>Range)
- ☑ P73i_is_translation_of(Subclass all)

APPENDIX 6: GROOVY XPATH EXTRACTION

Source Script	String userDir = System.getProperty("user.home");
	rootDir = userDir+"/TXM/scripts/user/";
	def TEI = new XmlSlurper().parse(rootDir+"000027_mod2.xml")
	
	TEI.text.body.div1.div2.p.note.p.persName.findAll { it.@ref.text()}.each { persName->
	    println persName.text()
	}

TEI	00027_mod2.xml
ELEMENT	persName
XQUERY	**findAll "persName", iterate "ref" not null**
XPATHS	**1. TEI.text.body.div1.div2.p.persName**
	2. TEI.text.body.div1.div2.div3.listBibl.bibl.hi.hi.persName
	3. TEI.text.body.div1.div2.p.note.p.persName

Result Path 1	Result Path 2	Result Path 3
Bffon	Ch.Bonnet	A.G.
Boerhaave	Ch.Bonnet	A.G.
Brisson	Derham	A.W.
Buch	Gianv.Petrini	Abbot
Cepede	Jos.Jac.	Ad.L.
Darwin	Karsten	Bligh
Hedwig	Leske	Blumhof
Ingen-Hou	Linn	C. A. S. Hoffmanns
Ingen-Hou	Marcell.Malpighii	C.E.
Klreuter	Nehem.Grew	C.Fr.
Klreuter	Ph.Andr.	C.Haidingers
Lc	Raab	Cicero
Lc	Steph.Hales	Emmerlings
Lichtenberg	Struve	Girtanner
Linn	Valm. de Bomare	Gnther
Linn	Vinc.Petagnae	Haller
Linn	Woodward	Haller
Linn		Herm.Sam.
Linn		Hildebrandt
Mhring		Hollmanns
Reaumur		J.Ellis
Spallanzani		J.F.
Swammerdam		Kant
Voigts		Kant
		Kant
		(next 26 rows omitted for brevity...)

APPENDIX 7: SUMMARY WORK LOG

Student Name	Christopher H. Johnson
Student ID	#3147338
University	BTU Cottbus-Senftenberg
Date	9 Jan. 2014

PROJECT CONTACTS

Name	Email	Initials: (Correspondence Log ID)
Christopher Johnson	<chjohnson39@gmail.com>	CHJ
Joerg Wettlaufer	<jwettla@gwdg.de>	JW
Professor Michael Schmidt	<dekanat4@tu-cottbus.de>	PMS
Christine Melchert	<umweltplanung@tu-cottbus.de>	UWS
Sree Ganesh Thotempudi	<sree-ganesh.thotempudi@gcdh.de>	SGT
Martin Scholz	<martin.scholz@informatik.uni-erlangen.de>	MS
Fabian Cremer	<fabian.cremer@sub.uni-goettingen.de>	FC
Frank Fischer	<frank.fischer@zentr.uni-goettingen.de>	FF
Katharina Stephan	<katharina.stephan1@gmail.com>	KS
Sanjeev Laha	<sanjeev.laha@stud.uni-goettingen.de>	SL
Andrea Schneider	<andrea.schneider@gcdh.de>	AS

CORRESPONDENCE AND ISSUES LOG

Seq	Type	Date	Sender	CC	Subject
1	Email	14 July 2013 16:21	**CHJ**		Request for Study Project to AS at Göttingen
2	Email	6 August 2013 13:13	**AS**		Introduction to GCDH, Referral to Jörg Wettlaufer
3	Email	Sat, Aug 17, 2013 at 3:20 PM	**CHJ**		Request for Study Project, Introduction of CHJ to JW
4	Email	Tue, Aug 20, 2013 at 11:30 AM	**JW**		Invitation for CHJ to participate in GCDH Project JW Defintion of Semantic Blumenbach: "Semantic Blumenbach" and aims to model the relationship between the publications of Johann Friedrich Blumenbach, a professor for anatomy and natural history in 18th century Goettingen, and the collection of objects that he initiated and fostered at Goettingen at the same time. For modelling this with RDF and using CIDOC (Erlangen CRM) we use the WissKI Environment that was developed at Erlangen and Nürnberg and make use of the material already digitized by the Academy project "Johann Friedrich Blumenbach - online"
5	Email	Tue, Aug 20, 2013 at 9:42 PM	**CHJ**		Arrange for first phone call on 21 August
6	Phone Call	21 August, 2013 at 8:00 PM	**JW**		Discussed nature of study project and CHJ interest. Arranged for meeting in Dresden Monday, 26 Aug at 18:00
7	Email	Wed, Aug 21, 2013 at 8:48 PM	**CHJ**		Terminbestätigung for Dresden meeting
8	Email	Thu, Aug 22, 2013 at 7:34 AM	**JW**		Confirmation of Meeting
9	Meeting	26 Aug. 2013	JW / CHJ		Disussed history of project, professional backgrounds, CHJ explained his understanding of semantic ontology and CIDOC. JW explained how Semantic Blumenbach fits into the U. Gottingen, DFG. GCDH frame.

10	Email, attachment	Sun, Sep 1, 2013 at 11:49 AM	**CHJ**		CHJ Draft Proposal for Study Project sent to JW and PS. Project Defined as: 1) Documentation of project framework, standards, reference models, and code base. 2) Evaluation of application tools, techniques and methods. 3) Problem-solving and development tasks as directed by Prof. Dr. Joerg Wettlaufer
11	Email	Mon, Sep 2, 2013 at 5:22 PM	**JW**		Approval of Proposal
12	Email, attachment	Mon, Sep 2, 2013 at 9:13 PM	**CHJ**		Final Proposal Sent to JW
13	Email, attachments	Wed, Sep 4, 2013 at 3:43 PM	**JW**		Resources provided: • Guest Login for Blumenbach-Online website @ http://dhfv-ent2.gcdh.de/blumenbach/semblu/blumenbach.php • Confidentiality Agreement (pdf) • Semantic Blumenbach 'Wiki' (pdf)
14	Email, attachment	Wed, Sep 4, 2013 at 11:08 PM	**CHJ**		Sent authorized confidentiality agreement
15	Email, attachments	Fri, Sep 6, 2013 at 11:16 AM	**JW**		Resources provided: • Login for Wisski environment @ http://dhfv-ent2.gcdh.de/blumenbach/wisski/ • NER to TEI Poster (pdf) Metadata Catalogues: • Metadaten für die Erfassung „naturhistorischer Blumenbachiana“: Erläuterungsliste (pdf) • Sample entry: Zoology Inventory Number 0343 (pdf)
16	Email, attachments	Mon, Sep 9, 2013 at 8:10 AM	**CHJ**		**Mapping ECRM – LIDO-** Response to JW Question about CIDOC *E84 Information Carrier* Resources Provided: • LIDO v.1 Specification (pdf) • Example of CIDOC Mapping using LIDO (pdf)
17	Email	Mon, Sep 9, 2013 at 8:53 AM	**JW**		Acknowledgement of CHJ response to question.

18	Email	Thu, Oct 10, 2013 at 10:07 AM	**CHJ**		Notification of Return from Urlaub by CHJ to JW
19	Email	Thu, Oct 10, 2013 at 10:13 AM	**JW**		Acknowledged CHJ status update. JW provided brief status update on Semantic Blumenbach: WissKI problem identified 'TEI files are too big' JW noted that Erlangen / Nürnberg has received new funding for continuation of WissKI development
20	Meeting	15.10.13	PMS		Confirmed and officially authorized registration of study project. Supervision by Prof. Dr. Michael Schmidt. (Lehrstuhl Umweltplanung)
21	Email	Wed, Oct 16, 2013 at 1:11 PM	**CHJ**		Notification to JW about confirmation of study project registration. CHJ indicates desire to use MediaWiki with protected namespace as a documentation platform. CHJ indicated his understanding of the problems with Drupal 6 as a 'legacy' application development environment and the need for WissKi to be moved to Drupal 7.
22	Email	Thu, Oct 17, 2013 at 1:16 PM	**JW**		JW acknowledged and approved use of MW as a documentation platform. JW asks if GCDH to Erlangen / Nuernberg German project correspondence would be useful.
23	Email	Sun, Oct 20, 2013 at 8:09 PM	**CHJ**		CHJ provided JW with login for documentation Wiki @ http://erbeinformatik.org/w/index.php?title=SP:Semantic_Blumenbach CHJ describes the documentation approach and current progress.
24	Email	Tue, Oct 22, 2013 at 10:54 PM	**JW**		JW acknowledged receipt of login information for documentation Wiki.
25	Email	Wed, Oct 23, 2013 at 3:04 PM	**CHJ**		CHJ noted study progress: • evaluation of XML-TEI format • complete review of NER methods • definition of project status and phases based on data model and task list • identified translating / triplifying the TEI to RDF as current main project objective CHJ notes a request for project specific TEI file (000027 1799 Blumenbach) and an XSLT stylesheet for rendering. CHJ notes his investigation of the DTA (Deutsches Textarchiv), an online repository of annotated text files. CHJ requests clarification on the part of the project that will model the Blumenbach Sammlungen objects

26	Email, attachments	Wed, Oct 23, 2013 at 5:38 PM	**JW**		Resources provided; • 000027.xml (Blumenbach TEI sample text) • wisski_triplify_neu.zip • triplify.xsl • Presenation about Semantic Blumenbach Project (pdf) JW clarifies that Sammlungen objects will be mapped to a CRM by others and imported into the WissKI system via OBDC..
27	Email	Fri, Oct 25, 2013 at 3:55 PM	**JW**		JW requests CHJ status. JW indicates that he has reviewed the MW documentation. JW points to the possibility of obtaining a stylesheet from Blumenbach-Online JW suggests the TEICHI module for Drupal 7 as a possible server based TEI display and reading solution.
28	Email	Fri, Oct 25, 2013 at 6:26 PM	**CHJ**		CHJ notes study progress: • evaluation of XML content management systems and readers with the intent of rendering the 000027.xml TEI file • Installed and evaluated Omeka. • Tested OAI-PMH Harvester • Reviewed MW based Transcribe-Bentham @ http://www.transcribe-bentham.da.ulcc.ac.uk/td/Transcribe_Bentham. • Installed and evaluated Drupal 7 with RDFx extension. • Evaluated GATE CHJ expressed that study focus is 'to get familiar with the latest developments and techniques specific to text engineering, annotation, and LOD'
29	Email	Sat, Oct 26, 2013 at 10:05 AM	**JW**		JW recommends the 'Archeo 18 Project at SUB Goettingen' as an XML-TEI presentation environment
30	Email, attachments	Tue, Oct 29, 2013 at 12:27 PM	**CHJ**		Resources provided: • TXM generated description file of a DTA sourced Blumenbach TEI CHJ noted study progress: • evaluation of TEICHI. Pointed out it is very simple and has limitations. • Used Oxygen to view and annotate 000027.xml schema and published it http://erbeinformatik.org/sp/sbschema.html.

					• Indicated evaluation of TXM o imported DTA sourced Blumenbach TEI o Explained an observed problem with importing the 000027 TEI file into TXM
31	Email	Thu, Oct 31, 2013 at 11:32 AM	**JW**		JW acknowledges and appreciates the schema documentation. JW indicates important elements are <term> <rs>, <persName>, <placeName>,<date>. <rs> is used to model the relationship between Objects and texts.
32	Email. attach ments	Mon, Nov 4, 2013 at 5:05 PM	**CHJ**		**TXM and 0027** Resources provided: TXM generated files: • Silber Progression Graph • Concordence example • Lexicon of the 000027 file. CHJ indicates progress: • that he fixed the problem with the TXM import • evaluation and study of TXM as a TEI development platform • developed an understanding of linguistic applications of XML-TEI for 'lemmatization' • developed an understanding of 'tokenization' in relation to the TXM search engine
33	Email	Tue, Nov 5, 2013 at 10:36 AM	**JW**		JW explains issue with importing the TEI into TXM. JW indicates his priorities are: • modelling the text - object relationships in the wisski system. • Updating the Application ontology in OWL • The Data Model needs to be transfered to the paths and the OWL Ontology. • Expanding the semblu Entities in ECRM
34	Email. attach ments	Tue, Nov 5, 2013 at 3:27 PM	**CHJ**		Resources provided: • Modified (debugged) version of the 000027.xml file that will import into TXM • Plain text output of the tree tagger that lemmatized all of the words in the TEI CHJ explains the import method he used for TXM. Speculates that the limitation in element XQuery function is the result of not having a custom XSLT to explain the elements to the software. Phone call is arranged for morning of Friday, Nov. 8
35	Email	Wed, Nov 6, 2013 at 12:26	**JW**		JW acknowledges the receipt of the files. Indicates that originally they had used GATE to generate XML output with linguistic annotation

		PM			Confirms phone call on Friday.
36	Email	Wed, Nov 6, 2013 at 5:42 PM	**CHJ**		CHJ reconfirms phone call.
37	Email	Fri, Nov 8, 2013 at 8:14 AM	**JW**		JW cancels phone call
38	Email	Sat, Nov 9, 2013 at 10:18 PM	PMS		CHJ sends brief update to Prof. Schmidt. Notes problem with accessing TextGrid since BTU is not part of the DFN-AAI. https://www.aai.dfn.de/ Requests Prof. Schmidt to investigate if DFN access is possible.
39	Email. attachment	Sun, Nov 10, 2013 at 7:55 AM	**JW**		Resources provided: • Blumenback_changedrc.rnc JW forwarded correspondence with developer of the NER parser, Ganesh. Noted an issue with new line break tags in the output file. <lbtype="inWord"/>
40	Email. attachment	Sun, Nov 10, 2013 at 7:58 AM	**JW**		Resources provided: 000027.xml output file that Ganesh sent to JW on 31 October
41	Email	Sun, Nov 10, 2013 at 8:11 AM	**JW**		JW explains difficult situation with NER development and developer.
42	Email	Mon, Nov 11, 2013 at 2:42 AM	**CHJ**		CHJ indicates progress: • Using TXM, Oxygen and TextGrid together • validated the latest 0027 file against the relax schema • corrected 0027 file by deleting occurrences of <lbtype="inWord"/> • debugged and reformed new 0027 for import into TXM • Used Groovy language to write scripts in TXM for XML value extraction • developed an understanding of XPATH and XQUERY • investigated using Saxon EE with XSLT to transform XSD to RDFS
43	Email	Mon, Nov 11,	**JW**		JW arranges phone call at 10:30.

		2013 at 10:02 AM			Suggests the Graphite PHP library @ http://graphite.ecs.soton.ac.uk/ as a possible XML to RDF transformation tool
44	Email	Mon, Nov 11, 2013 at 10:36 AM	**CHJ**		CHJ confirms availability for phone call.
45	Skype	Mon, Nov 11, 2013 at 10:45 AM	**JW-CHJ**		Lengthy discussion of project status.
46	Email. attachments	Wed, Nov 13, 2013 at 3:23 PM	**JW**		Resources provided • SQL Dump of NER reference entity values: • extractor.php • Sammlungsobjekte_Stichworte.txt JW describes these files as: • lists which uses the Java-parser to tag the TEI files • a quite basic script to extract the names of the object • resulting list of strings - Database ID in a text file
47	Email. attachments	Fri, Nov 15, 2013 at 11:24 AM	**CHJ**		CHJ indicates progress: • Developed Groovy script for extracting XML value data from TEI in TXM. • Noted the importance of consistent or simplified TEI nodal structure in XPATH definitions o noted that the 000027 TEI file has many path variations for identically described elements. • Resources provided: • XML-TEI Extraction Results.xslx This is a spreadsheet of values extracted from the <persName>, <placeName> and <term> elements using 3 different XPATHs for each element.
48	Email	Fri, Nov 15, 2013 at 5:59 PM	**JW**	FC MS	JW indicates that he believes XSLT extraction to be preferred over scripting because it does not require XPATH identification. JW suggests the SIG resource for TEI to CRM mapping
49	Email	Sun, Nov 17, 2013 at 7:45 PM	**CHJ**	FC MS	CHJ indicates his understanding of the SIG document. He notes that having XML:IDs may be essential for mapping a TEI XSD to RDFS. CHJ discusses the possible advantage of script value extraction over XSLT. CHJ asks if JW has an ODD file for the Blumenbach TEI schema.

50	Email	Mon, Nov 18, 2013 at 10:57 AM	**JW**		JW shares the Semantic Blumenbach Drop Box folder. Resources provided: • 000027_facs_jpg (images of every page in jpg) • 000027_single_pages (in TEI Format • perl-scripts from the DTA
51	Email. attachments	Mon, Nov 18, 2013 at 7:27 PM	**CHJ**	FC MS	Resources provided: • Blumenbach TEI Relax Scheme based on 2.5 TEI Specification • ODD File generated with the Roma tool for the 000027 TEI schema CHJ inotes issues with current 000027 encoding as validated against the 2.5 specification CHJ explained how these files were created and their possible use. CHJ requests JW opinion on XML:IDs
52	Email. attachments	Mon, Nov 18, 2013 at 9:27 PM	**JW**		Resource Provided: • Document 'Redaktionelle TEI-Auszeichnung von Blumenbachvolltexten' JW indicates that there are compatibility constraints with Blumenbach Online that may limit the adaptation of the TEI schema (I.e. the addition of XML:IDs) JW notes the continued problem with the addition of the unwanted -<lb type="InWord"\> that occurs in semantically tagged words.
53		Mon, Nov 18, 2013 at 9:30 PM	**JW**		Forwarded email (Tue, 15 Oct 2013) from DTA to Blumenbach team on Blumenbach TEI schema
54		Mon, Nov 18, 2013 at 9:31 PM	**JW**		Forwarded email thread. Regarding the problem with unwanted -<lb type="InWord"\>
55	Email. attachment	Mon, Nov 18, 2013 at 9:32 PM	**JW**		Resource Provided: • Documentation and Results.zip (contains 000027.xml results file and a non-structural documentation of the schema)
56		Mon, Nov 18, 2013 at 10:17 PM	**JW**		Notification of Free Webinar on Intro to Semantic Web
57		Wed, Nov 20, 2013 at 2:37	**JW**		Notification of SWIB13 Workshop: Linked Data Publication with Drupal

		PM			
58		Wed, Nov 20, 2013 at 5:13 PM	**JW to SL**	CHJ	Forwarded email regarding continued investigation of -<lb type="InWord"\> issue
59		Fri, Nov 22, 2013 at 4:13 PM	**CHJ**		Status update. CHJ indicates the need for a short term progress report.
60	Email. attachments	25 November 2013 09:06	**JW**		Resources provided: • text-object relation diagram (png) • semblu OWL ontology • semblu OWL Protege project file JW suggests looking at modeling the linguistic object JW also notes that modeled entities (term, rs, place and person) should be included in OWL for facility of TEI import into WissKI
61	Email. attachments	26 November 2013 00:42	**CHJ**		CHJ acknowledges task request for modeling Linguistic Object. CHJ sends status report. Indicates has obtained a new laptop with Ubuntu, enabling use of new tools: • Protege is able to load the R libraries for the OwlViz module, (graphs for the modeling) • XQuery app called BaseX. (two sample query outputs: 'rs' and 'placeName') Resources provided: • Independent Study Project Report 25-11-13.pdf • rs query from BaseX • placeName query from BaseX
62		27 November 2013 16:51	**JW**		JW acknowledges receipt of BaseX sample outputs. Requests file format claritication.
63	Email. attachments	28 November 2013 15:26	JW		Fowarded email from Fabian Cremer ' Semblu ERM& OWL‘ that includes mapping TEI elements with UML tool. Resources: • semblu_erm.jpg • semblu_ontology.owl • semblu.object.violet.html

64	Email. attachments	28 November 2013 15:32	CHJ		CHJ shares Linguistic Object modelling diagram from Protégé. Notes and includes primary reference from the EAD centered paper “The Semantic Mapping of Archival Metadata to the CIDOC CRM Ontology”. CHJ notes that the model reflects the “4 top semantic hierarchies of an archive document” Resources provided: • document class top order hierarchy.png • Archival Metadata in CIDOC.pdf
65	Email. attachments	6 December 2013 12:02	**JW**	CHJ	CC response email to Fabian Cremer ' Semblu ERM& OWL‘ includes markup of UML diagram Resources provided: • Marked up UML graph • Datenmodell_Wisski.pdf
66	Email. attachment	8 December 2013 11:00	**JW**	FC	Detailed response about project status o JW notes that the correspondence log is very useful and detailed. o Indicates the the BaseX query outputs reveal some issues with the completeness of the Getty IDs, but the rs looked right. o Suggests that Oxygen and BaseX provide similar functionality o Likes the Protégé Graph, thinks that it may be useful for presentation and publication Discusses the status of the NER software development: • Fixed the problems with valid XML (release of 0.9 beta). (Attached the last result file). • Pattern tagging for the animals works (the output is valid XML) • the placenames and personnames tagging is still case insensitive o After this is fixed, correct the errors and also add some more tags for rs manually. o cross-check manually if there are any more possible references to the collection objects in the text. (800 object, this should be possible, all the more as we have already lists from the index of the book #27 (provided by Blumenbach-online) with page references.) o Uncertain whether to extract the entities from the text again and run the parser with updated lists • Primary goal for the NER should be to have the 12 German Editions of the "Handbuch der Naturgeschichte" ready by the end of January for ingestion into the WissKI system. Resources provided: o 000027.zip
67	Email	9 December 2013 11:54	**JW**	FC MS	JW indicates that a blocking bug in WissKI has been fixed: Part of the TEI-Modul (wisski-textmod modul)

68	Email	9 December 2013 15:07	**JW**	FC MS	JW provides feedback on the EAD model. Questions on how Authority Documents work, how to model structural TEI elements as part of the Document class. JW indicates that his goal is to model TEI elements and their relationship to the tagged entities.
69	Email	10 December 2013 09:52	**JW**	CHJ	CC reply to Martin Scholtz discusses general issues with the WissKI system.
70	Email	10 December 2013 13:03	**CHJ**		Subject: Modelling TEI in CIDOC CHJ responds to detail questions from JW: 1. The relationship between the Document and P106 is circular and has no constraint so it is possible to have as many transitive 'child' Document objects as required. 2. 2. The authority document would be the TEI version schema declared as Types (XSD / RNG and ODD). This is important because it specifically validates the document annotations. CHJ indicates intention to model TEI elements within the 4 class EAD hierarchy
71	Email	13 December 2013 14:31	**JW**	CHJ FF KS	CC Email to Ganesh from JW Discusses specific tagging issues from the NER software development
72	Email	13 December 2013 16:13	**SGT**	FF KS	CC on Ganesh’s reply to JW about NER specifics
73	Email	17 December 2013 10:51	**CHJ**		CHJ indicates he has understood how to map TEI elements within the EAD hierarchy. Discusses the use of the Appellation Class to map to TEI elements. Discusses how Authority Documents could map to both the object and the TEI markup. References and includes paper: o WP5-T5_5-ead2crm-mapping-060728v0_2-final.doc
74	Email	19 December 2013 17:43	**JW**	MS FC	JW acknowledges validity of the use of Appellations. Includes detailed status quo mapping in the WissKI system.
75	Email	22 December 2013 12:12	**JW**	MS FC	JW links to current discussions about CIDOC modelling. http://www.ontotext.com/CRMEX http://www.youtube.com/watch?v=Ai7uhtRF7HM Indicates a new tool called Research Space http://www.researchspace.org/ JW poses specific questions about the context of Linguistic Object and Authority Documents in the hierarchy.

76	Email	25 December 2013 16:18	**CHJ**	MS FC	CHJ answers the questions posed by JW. Gives specific paths that model the Linguistic Objects, Authority Documents and Identifier CHJ discusses inheritance and the idea of relative subjects. Gives example of a possible mapping for <term> CHJ provides a link to the Arches project: http://archesproject.org/
77	Email	28 December 2013 16:58	**CHJ**	MS FC	CHJ indicates that WissKI models ontology paths from a bottom up approach and that this creates problems for inheritance. CHJ suggests the possibility of modelling using a 2 dimensional (x,y) approach rather than 1 (y)
78	Email	30 December 2013 22:26	**JW**	MS FC	JW acknowledges CHJ input on modelling.
79	Email	4 January 2014 15:33	**JW**	MS FC	JW notes: 'If we can use the pathbuilder, then we have to model the hierarchy of XML using the CIDOC hierarchy with eventually some additions in the application ontology layer. We have classes and subclasses in CIDOC and we can declare our own subclasses in the application Ontology (ontology.owl). The triples constitute a flat structure and hierarchy only comes by the inbuilt classes and subclasses in the CIDOC. So we do not have to model the CIDOC with the pathbuilder again but only put our data (terms, placenames, etc.) at the right place in the CIDOC and then make use of the inherent structure of the CIDOC.' JW indicates that the NER project phase is done. JW suggests a Skype call on Wednesday, 8 Jan.
80	Email	5 January 2014 16:44	**CHJ**	MS FC	CHJ asks if the project intends to produce an RDF file with only with the NER triples. Indicates that he was under the assumption that the product would be the complete document. CHJ confirms that the ontology will not be part of the output file. Says 'the challenge and problem is to define the 'flat' triples so that they are coherent with single entities. CIDOC does not really make this easy, because sometimes is takes at least two properties (like P1 and P2) to fully describe an entity attribute, and unless the value is repeated for each domain, the link between the intermediate class may get lost' CHJ references the Arches data model where a mapping steps table is used to connect triples together. CHJ points out that cardinality is difficult to enforce in a flat data model. CHJ confirms Skype call on Wednesday.
81	Email	6 January 2014 21:50	**JW**	MS FC	JW clarifies that the current project output goal is to produce triples in the ARC2 triple store of the Drupal 6 installation. JW says that the text and the triples exist separately in Drupal. The text is related to the triples with automatically

					assigned <div> IDs. JW also points out that 'Cardinality is an issue with the reference string, as we have a List of Database IDs in this tag which have to be converted into a Turtle list' JW confirms Skype call on Wednesday at 10:00
82	Email	7 January 2014 22:09	**CHJ**		CHJ confirms Skype call for 8 Jan at 10:00
83	Skype	8 January 2014 10:00	**Skype call: JW / CHJ**		Discussion includes: Immediate project objectives and modelling goals. CHJ logs into WissKI and looks at paths and asks JW direct questions about existing path objects. CHJ points out that the use of the Note class for value strings is incorrect. CHJ recommends using E55 Type as an intermediate mapping step whenever an E42 Identifier is used and suggests also including a reference to the Authority File for the Identifier definition. CHJ says that he would model these paths in Protégé and send JW graphs. Both JW and CHJ feel that the modelling approach needs careful consideration. JW expresses that an experimentation approach may yield more answers. CHJ points out that connecting the TextGrid object to the WissKI Drupal text and triples is not possible without establishing a common identifier for the div objects. JW agrees, but indicates that this is not a current project concern. CHJ indicates a desire to visit Göttingen in February and JW acknowledges that this may be possible. CHJ indicates that he needs to produce the study project output paper by the middle of February. JW indicates that he would be happy to take a look at it.
84	Email	9 January 2014 15:50	**CHJ**		CHJ sends a copy of the summary correspondence log to Prof. Michael Schmidt
85	Email	10 January 2014 09:57	**UWS**	SC	Umweltplanung Sekretariat responds to summary log. Indicates that final paper should be sent to <umweltplanung@tu-cottbus.de>
86	Email	10 January 2014 08:19	**CHJ**		CHJ thanks JW for Skype call. Indicates that he is looking at WissKI pathbuilder code in order to understand mapping.
87	Email	12 January 2014 19:32	**CHJ**		CHJ tells JW that he has looked at ARC2 code and found an explanation for the pathbuilder syntax. He also reveals that the authority file mapping problem could be solved by including a full URI path for any identifiers that are validated by non-local authorities.
88	Email	12 January 2014 22:29	**JW**	MS FC	JW acknowledges that using full URIs for IDs enhances the possibilities for LOD.
89	Email	13 January	**MS**	JW	Martin Scholz provides detailed and clear explanations on how WissKI pathbuilder works. Indicates that he would

		2014 12:24		FC	be able to offer assistance if WissKI internals need to be extended.
90		19 January 2014 22:35	**CHJ**	JW MS FC	CHJ thanks Martin Scholz for the detailed explanations. Offers his thoughts about rules and OWL DL in relation to a presentation by Markus Krötzsch.

RESOURCE LIST

FROM JW	**FROM CHJ**
• Guest Login for Blumenbach-Online website @ http://dhfv-ent2.gcdh.de/blumenbach/semblu/blumenbach.php • Confidentiality Agreement (pdf) • Semantic Blumenbach 'Wiki' (pdf) • Login for Wisski environment @ http://dhfv-ent2.gcdh.de/blumenbach/wisski/ • NER to TEI Poster (pdf) • Metadaten für die Erfassung „naturhistorischer Blumenbachiana“: Erläuterungsliste (pdf) • Sample entry: Zoology Inventory Number 0343 (pdf) • 000027.xml (Blumenbach TEI sample text) • wisski_triplify_neu.zip • triplify.xsl • Presentation about Semantic Blumenbach Project (pdf) • Blumenback_changedrc.rnc • 000027.xml output file that Ganesh sent to JW on 31 October • SQL Dump of NER reference entity values: • extractor.php • Sammlungsobjekte_Stichworte.txt • 000027_facs_jpg (images of every page in jpg) • 000027_single_pages (in TEI Format • perl-scripts from the DTA • Document 'Redaktionelle TEI-Auszeichnung von Blumenbachvolltexten' • Documentation and Results.zip (contains 000027.xml results file and a non-	• LIDO v.1 Specification (pdf) • Example of CIDOC Mapping using LIDO (pdf) • Description file of a DTA sourced Blumenbach TEI from TXM • Silber Progression Graph from TXM • Concordence example from TXM • Lexicon of the 000027 file from TXM • Modified (debugged) version of the 000027.xml file that will import into TXM • Plain text output of the tree tagger that lemmatized all of the words in the TEI • XML-TEI Extraction Results.xslx • Blumenbach TEI Relax Scheme based on 2.5 TEI Specification • ODD File generated with the Roma tool for the 000027 TEI schema • Independent Study Project Report 25-11-13.pdf • rs query from BaseX • placeName query from BaseX • document class top order hierarchy.png • Archival Metadata in CIDOC.pdf • WP5-T5_5-ead2crm-mapping-060728v0_2-final.doc

structural documentation of the schema) from Ganesh • text-object relation diagram (png) • semblu OWL ontology • semblu OWL Protege project file • semblu_erm.jpg • semblu_ontology.owl • semblu.object.violet.html • Marked up UML graph • Datenmodell_Wisski.pdf • 000027.zip	

PRIMARY IN-PROCESS TASKS:

1	Documentation of *Semantic Web* Web Terms and Concepts
2	Analysis of Blumenbach Project Integration Issues with other Projects (e.g. WissKI, Blumenbach-Online and DTA))
3	Evaluation of *Semantic Web* Tools (Content Management (Drupal (TEICHI, RDFx), Omeka, MediaWiki), XML-TEI Authoring (Oxygen, TXM, TextGrid, GATE), Languages and Scripts (Perl. Ruby, Python, Groovy), Ontology Editors (Protege), Xquery Extraction and Node Graphing (BaseX XML Database).
4	Review of Code Base (GCDH NER Parser and entity sources (from Stichworte and SQL dump), 0027.xml TEI format)
5	Problem solving
6	TEI (Debugging, validation, schema evaluation with Roma)
7	XQUERY and XPATH
8	Tokenization, Concordance, Lexical Analysis and Lemmatization (SOLR, Saxon, TXM Groovy) and XSLT Transformations
9	Modeling and Ontology (ECRM OWL in WissKI / Protege)

BIBLIOGRAPHY

ARVIDSSON, FREDRIK AND ANNIKA FLYCHT-ERIKSSON. (n.d). ONTOLOGIES. [pdf]. Available online: <http://www.ida.liu.se/~janma/SemWeb/Slides/ontologies1.pdf>

BAKER, THOMAS. (2000). A Grammar of Dublin Core. Available online: <http://www.dlib.org/dlib/october00/baker/10baker.html>

BOUNTOURI, LINA AND MANOLIS GERGATSOULIS. (2012). The Semantic Mapping of Archival Metadata to the CIDOC CRM Ontology. [pdf]. Available online: <http://www.tandfonline.com/doi/pdf/10.1080/15332748.2011.650124>

CARRASCO, LAIS BARBUDO. (2013). Information Integration: Mapping Cultural Heritage Metadata into CIDOC CRM. [pdf]. Available online: <http://portal.febab.org.br/anais/article/download/1409/1410>

COBURN, LIGHT, MCKENNA, STEIN, AND VITZTHUM. (2010). LIDO - Lightweight Information Describing Objects Version 1.0. [pdf]. Available online: <http://www.lido-schema.org/schema/v1.0/lido-v1.0-specification.pdf>

CROFTS, NICK. (2003). MDA Spectrum CIDOC CRM mapping. [pdf]. Available Online: <http://www.cidoc-crm.org/docs/MDA%20Spectrum_CIDOC_CRM_mapping.pdf>

DOERR, MARTIN. (1998). Data Example of the CIDOC Reference Model - Epitaphios GE34604 –. [pdf]. Available Online: <http://www.cidoc-crm.org/docs/crm_example_1.pdf>

DOERR, MARTIN. (2002). Mapping a Data Structure to the CIDOC Conceptual Reference Model. [ppt] Available Online: < http://www.cidoc-crm.org/crm_mappings.html>

ENGLISH HERITAGE. (2012). MIDAS Heritage: The UK Historic Environment Data Standard. [pdf]. Available online: < http://www.english-heritage.org.uk/publications/midas-heritage/midas-heritage-2012-v1_1.pdf>

GCDH. (2014). Akademie Der Wissenschaften Zu Göttingen (ADW). Available Online <http://www.gcdh.de/en/projects/tp1-dlvm/adw>

GERBER, VAN DER MERWE AND BARNARD. (2007). A Functional Semantic Web Architecture. [pdf]. Available online: <http://ksg.meraka.org.za/~agerber/Paper152.pdf>

GRUBER, THOMAS. (1993). A Translation Approach to Portable Ontology Specifications. [pdf]. Available online: <http://tomgruber.org/writing/ontolingua-kaj-1993.pdf>

GRÜN, KRAMIS, HOLUPIREK, ET. AL. (2006). Pushing XPath accelerator to its limits [pdf] Available online: <http://kops.ub.uni-konstanz.de/bitstream/handle/urn:nbn:de:bsz:352-opus-23294/push_accel_expdb06.pdf?sequence=1>

HEIDEN, SERGE. (2013). Exploiting TEI-annotated data with TXM. [pdf] Available online: <http://digilab2.let.uniroma1.it/teiconf2013/wp-content/uploads/2013/09/Heiden.pdf>

HITZLER, KRÖTZSCH, AND RUDOLPH. (2009). Knowledge Representation for the Semantic Web Part II: Rules for OWL. [pdf] Available online: <http://www.semantic-web-book.org/w/images/5/5e/KI09-OWL-Rules-2.pdf>

HORRIDGE, MATTHEW. (2011). A Practical Guide To Building OWL Ontologies Using Protege 4 and CO-ODE Tools Edition 1.3. [pdf] Available online: <http://130.88.198.11/tutorials/protegeowltutorial/resources/ProtegeOWLTutorialP4_v1_3.pdf>

KROKE, CLAUDIA. (2010). Johann Friedrich Blumenbach Bibliographie seiner Schriften. Universitätsverlag Göttingen. [pdf] Available online: <http://rep.adw-goe.de/bitstream/handle/11858/00-001S-0000-0001-CC58-7/Kroke_Bibliographie_PDF.pdf?sequence=1>

LE BOEUF, DOERR, ORE, AND STEAD. (2013). Definition of the CIDOC Conceptual Reference Model Version 5.1. OWL 2 Web Ontology Language Primer (Second Edition.[pdf] Available online: <http://www.w3.org/TR/2012/REC-owl2-primer-20121211/>

PITZALIS, NICCOLUCCI, ET. AL. (2010). LIDO and CRMdig from a 3D Cultural Heritage Documentation Perspective. [pdf]. Available online: <http://www.athenaeurope.org/getFile.php?id=685>

SUB. (2014). Johann Friedrich Blumenbach – online - Project details. Available Online. <http://www.sub.uni-goettingen.de/en/projects-research/project-details/projekt/johann-friedrich-blumenbach-online/>

STASINOPOULOU, DOERR, PAPATHEODOROU, AND KAKALI. (2007). EAD mapping to CIDOC/CRM. [pdf]. Available online: <http://www.cidoc-crm.org/workshops/finland_helsinki_20102801/N13_28Jan2010%20Christos%20Papatheodorou.pdf>

UNIVERSITY OF MANCHESTER. (2005). Ontology Reasoning:Why do We Want It? [pdf]. Available online: <http://www.computational-logic.org/content/events/iccl-ss-2005/lectures/horrocks/part3a-reasoning.pdf>

WETTLAUFER AND THOTEMPUDI. (2013). Named Entity Recognition in Historical Texts from the Natural History Domain. [pdf]. Available online: <http://www.gcdh.de/files/2013/6429/9184/Wettlaufer_Thotempudi_2013_NER_final.pdf>

WETTLAUFER, THOTEMPUDI AND CREMER. (2012). Workshop: Semantic Web Applications in the Humanities. [pdf]. Available online: <http://www.gcdh.de/files/4613/5548/7941/Einfuehrung_Workshop_Semantic_Web_Applications_2012.pdf>

FIGURE REFERENCES

Figure 1: Semantic Blumenbach Structure	WETTLAUFER, ET. AL. (2012).
Figure 2: Semantic Web Model.	GERBER, VAN DER MERWE AND BARNARD. (2007).
Figure 3: An RDF Triple	BAKER, THOMAS. (2000).
Figure 4: TBox and Abox Axioms	UNIVERSITY OF MANCHESTER. (2005).
Figure 5: WissKI Pathbuilder Screenshot	GOERZ, SCHOLZ, FICHTNER, ET. AL. (2014).
Figure 6: Term Alabaster as a WissKI Triple	GOERZ, SCHOLZ, FICHTNER, ET. AL. (2014).
Figure 7: Mulatte Resource	BLUMENBACH ONLINE. (2014).
Figure 8: WissKI Triple Presentation	GOERZ, SCHOLZ, FICHTNER, ET. AL. (2014).