Schriften des Instituts für Dokumentologie und Editorik — Band 2

# Kodikologie und Paläographie im digitalen Zeitalter

---

# Codicology and Palaeography in the Digital Age

herausgegeben von | edited by

Malte Rehbein, Patrick Sahle, Torsten Schaßan

unter Mitarbeit von | in collaboration with

Bernhard Assmann, Franz Fischer, Christiane Fritze

2009
BoD, Norderstedt

**Bibliografische Information der Deutschen Nationalbibliothek**:
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über **http://dnb.d-nb.de/** abrufbar.

Leicht veränderte Fassung für die digitale Publikation (siehe Vorwort).

---

Slightly modified version to be published digitally (see preface).

Publication réalisée avec le soutien d'Apices
Association Paléographique Internationale
Culture – Écriture – Société
Dotation J.M.M. Hermans.
http://www.palaeographia.org/apices/



Herstellung und Verlag: Books on Demand GmbH, Norderstedt
ISBN: 978-3-8370-9842-6
Einbandgestaltung: Katharina Weber
Satz: XeTeX und Bernhard Assmann

# Representation and Encoding of Heterogeneous Data in a Web Based Research Environment for Manuscript and Textual Studies

Daniel Deckers, Lutz Koch, Cristina Vertan

## Abstract

This paper describes the general architecture of a digital research environment for manuscript and textual studies (particularly those pertaining to ancient Greek and Byzantine texts), and it discusses some questions of data representation and encoding in the framework of such an online research platform. The platform is being developed by the project *Teuchos. Zentrum für Handschriften- und Textforschung*, established in 2007 by the *Institut für Griechische und Lateinische Philologie* (Universität Hamburg) in cooperation with the *Aristoteles-Archiv* (Freie Universität Berlin). Teuchos is a long-term infrastructural project of the *Universität Hamburg*. It is currently in its three-year initial phase which is being co-funded by the German Research Foundation (DFG) through the "Thematic Information Networks" scheme within the "Scientific Library Services and Information Systems" programme. We introduce the main object types to be handled by our system and describe the overall functionality of the online platform. The paper focuses on the representations of two main object types: manuscripts as textual witnesses and watermarks, with an emphasis on the former. Since the adequate encoding of different layers of structure of a transmitted text is particularly relevant to optimising users' choices of navigating both digital images of the containing manuscripts and trancriptions of the text contained, this topic is discussed in some detail. We introduce the formal data model and the corresponding encoding for the object types discussed. The project encodes textual data in XML, aiming for TEI conformance where possible. Since no accepted XML model exists for the encoding of metadata within a watermark collection, we briefly explain how we chose to model the objects to accomodate the collections the project is making accessible.

## Zusammenfassung

Der folgende Aufsatz beschreibt die Gesamtarchitektur einer digitalen Arbeitsumgebung für Handschriften- und Textforschung (insbesondere im Bereich altgriechischer und byzantinischer Texte) und Lösungsansätze zu einigen Problemen der Datenrepräsentation und Kodierung im Rahmen einer solchen Online-Plattform. Die Plattform

*Kodikologie und Paläographie im Digitalen Zeitalter – Codicology and Palaeography in the Digital Age.* Hrsg. Malte Rehbein, Patrick Sahle und Torsten Schaßan, unter Mitarbeit von Bernhard Assmann, Franz Fischer und Christiane Fritze. Schriften des Instituts für Dokumentologie und Editorik 2. Norderstedt: Books on Demand, 2009. 79–95.

wird durch das 2007 am *Institut für Griechische und Lateinische Philologie* (Universität Hamburg) gegründete Projekt *Teuchos. Zentrum für Handschriften- und Textforschung* in Kooperation mit dem *Aristoteles-Archiv* (Freie Universität Berlin) entwickelt. Teuchos ist als langfristige Infrastruktureinrichtung der *Universität Hamburg* angelegt und befindet sich derzeit in seiner dreijährigen Startphase, die von der Deutschen Forschungsgemeinschaft (DFG) im Programm »Themenorientierte Informationsnetze« des Förderinstruments »Wissenschaftliche Literaturversorgungs- und Informationssysteme (LIS)« durch eine Anschubfinanzierung mitgetragen wird. Wir stellen zunächst die wichtigsten Arten von Objekten vor, die das System verwendet, um dann die übergreifende Funktionalität der Plattform zu beschreiben. Der Schwerpunkt liegt hier auf der Darstellung zweier zentraler Objektarten: Handschriften in ihrer Funktion als Textzeugen sowie Wasserzeichen, wobei die Handschriften ausführlicher behandelt werden. Insbesondere wird auf ein geeignetes Kodierungsmodell für verschiedenartige Strukturierungsebenen handschriftlich überlieferter Texte eingegangen, da ein solches von zentraler Bedeutung ist, um den Nutzern möglichst vielseitige Möglichkeiten zu bieten, einerseits durch die Digitalisate der texttragenden Handschriften, andererseits auch durch die Transkriptionen der enthaltenen Texte zu navigieren. Formale Datenmodelle und die zugehörige Kodierung für die behandelten Objektarten werden kurz dargestellt. Innerhalb des Teuchos-Projekts werden Textdaten in XML kodiert, soweit möglich TEI-konform; da kein etabliertes XML-Modell für die Kodierung von Metadaten innerhalb einer Wasserzeichensammlung existiert, wird zudem die Objektmodellierung für die durch das Projekt online bereitgestellten Sammlungen skizzenhaft erläutert.

## 1 Introduction

This paper briefly describes the general architecture of a digital research environment for manuscript and textual studies, as well as discussing some questions of data representation and encoding. The project *Teuchos. Zentrum für Handschriften- und Textforschung* was initiated in 2007 by the *Institut für Griechische und Lateinische Philologie (Universität Hamburg)* in cooperation with the *Aristoteles-Archiv (Freie Universität Berlin).* Teuchos is a long-term infrastructural project of the *Universität Hamburg* that is currently in its three-year initial phase which is being co-funded by the German Research Foundation (DFG) through the "Thematic Information Networks" scheme within the "Scientific Library Services and Information Systems" programme.

In its final form Teuchos is to provide a web based research environment suited for manuscript and textual studies, offering tools for capturing, exchange and collaborative editing of primary philogical data. The data shall be made accessible to the scholarly

community as primary or raw data in order to be reusable as source material for various individual or collaborative research projects. This objective entails an open access policy using creative commons licenses regarding the content generated and published by means of the platform (esp. digital images of manuscripts may have to be handled restrictively dependant upon the holding institutions' policies). The software developed in the course of the project will be made available under free open source licensing as a contribution to the evolving diversity of digital humanities tools and applications.

Distinctive features of the Teuchos platform are the integration of heterogeneous research data (cf. 2) and the participation of different user groups in the generation and enhancement of the content. The system as a whole is geared to the needs and preferences of specialised research (rather than to the presentation of library treasures to a wider public). The following use cases are forseen:

- Provision of data facilitating the use of digitised manuscripts (created and shared by different user groups), ranging from structural information regarding the intellectual content of the manuscript to transcriptions containing indications of variant readings.
- Provision of digitised manuscripts accompanied by (partial) transcriptions both as a basis for further editorial work and to make core information on the content and the manuscript tradition available (and citable) to the scholarly community at the same early stage. While in certain cases this first step may be the only one that will be taken, with a view to the other cases it may also be considered a methodological improvement for textual studies in general to render the separate stages of the editorial process verifiable.
- Collaboration of networked researchers independent of time and space as a prerequisite for the analysis and utilisation of special materials, e.g. domain-specific texts and inaccessible or damaged sources such as palimpsests.
- An evolving collection of manuscript descriptions gives access to detailed information on codicology, manuscript history and textual transmission. This material derives from autoptical library studies and is thus often inherently sporadic and disjointed; on the other hand the collection is independent of library catalogisation projects and open to the collaboration of researchers worldwide who contribute according to their respective field of expertise and/or their serendipitous findings.
- A flexible model allows for the integration of manuscript descriptions of varying depth. A substantial amount of material taken from both published and unpublished materials of the *Aristoteles Graecus* offers a model for comprehensive and highly structured descriptions. The complex relationships between manuscripts belonging to a coherent textual tradition, e.g. common sources, scribes, owners, annotators etc. offer multiple possibilities of inter-linkage, which may be used for the individual exploration and eventually for automated analysis of such a corpus.

- Value is continuously added by way of cross-linking and data exchange with existing online resources relevant to manuscript and textual studies, especially library catalogues and manuscript bibliographies, but also archives of texts and digital libraries.

A prerequisite for the intended uses of the Teuchos platform are open data formats conformant to established community standards, the most important for several areas being TEI P5. Standards and data models to be used were also chosen with a view to the later addition of mechanisms to facilitate interoperability on the level of both basic metadata (e.g. through support of OAI-PMH) and of complex data sets (e.g. automatic conversion of the larger part of manuscript description data to the Manuscriptum XML format).

Following an overview of the system in general, we will proceed to discuss some specific encoding issues arising from the combination of materials we provide, and which we believe to be somewhat particular.

## 2 Teuchos Functionality and its Use for Codicology and Palaeography

The system we describe is meant to offer scholars and editors in the field of Greek codicology and palaeography a powerful tool for creating, publishing and augmenting relevant research materials of the kinds described below, as well as offering extensive search options for the larger scholarly community interested in consulting such materials. An overview of the basic system architecture is given in Figure 1.

All our objects are stored in a Fedora repository. This technology was chosen both for reasons such as standards conformance, robustness, openness etc., and because it has a wide user base in the eSciences and eHumanities.

The user interacts with this repository via a web application that manages the editing, searching, and uploading processes. We have three general categories of users:

- System administrators with full access to all parts of the system. They exercise overall control of the content made available by the other users.
- Registered users have access to released material on the system, and in addition they can contribute by uploading materials, writing forum entries and accessing and editing both released and preliminary materials within the scope of their respective user group. We envisage a hierarchy of such users having different access types to various parts of the system.
- Public users, who can only view released materials that carry no additional restrictions.

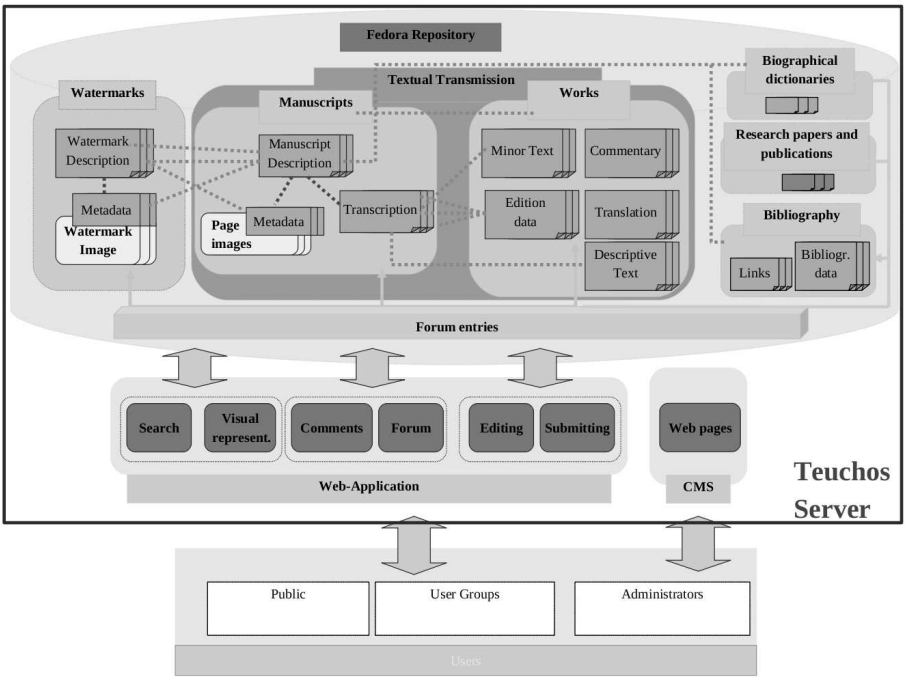


Figure 1. Teuchos Architecture.

There are several groups of digital objects to be stored in the Fedora repository:

- We store tracings of *watermarks* from dated paper manuscripts as digital images on the one hand, and descriptive data on these watermarks and their motif groups in an XML format on the other. Images are associated with Dublin Core (DC) and additional similarly structured metadata and linked to the descriptive data. A more detailed explanation is provided in 3.3.
- The *textual transmission* group is divided into two subgroups that are themselves subdivided: material related to individual *manuscripts* and material related to a particular *work*, e.g. a particular source text by a particular author.
- The *manuscript* group encompasses digital *page images* of manuscripts (or parts of manuscripts) that are aggregated on a per manuscript basis (cf. 3.1), scholarly *manuscript descriptions* that may reference page images if available for the manuscript described, and *transcription* data, which may range from a first set of basic structural data (cf. 3.2) to full transcriptions, and usually links to pages of ex-

actly one manuscript (exceptions are e.g. texts spanning more than one manuscript volume and re- or misbound manuscripts).
- The group of *works* encompasses a wide range of materials refering to a source text with its entire set of manuscripts rather than to one particular witness, and ranges from *full critical editions* (with several intermediate stages) and *translations* to various kinds of *commentaries* (and other explanatory or *descriptive* materials).
- The *biographical dictionary* group hosts prosopographic information on historical persons relevant to our other research materials.
- A special group is dedicated to *research papers* that may reference material from the other groups, without themselves falling into any of the other categories.
- Finally, *bibliographic data* (also including online resources) pertinent to specific research areas is collected and made available as a separate group of objects.

The Fedora repository offers a flexible mechanism to store heterogeneous information for one and the same object. A digital Fedora object consists of one unique identifier and several data streams. This facilitates the handling of these objects as basic relationships between objects and basic DC data for each object stored in separate streams from the main XML or image data as described above. More detailed information on individual authorship and responsibility for any particular part of one textual object produced cooperatively by a group of scholars, which is annotated upon data entry, can thus also be stored separately in an additional data stream. Should later extensions to the project result in additional kinds of annotation (e.g. linguistic or semantic), these could also easily be added as stand-off markup in separate streams without repercussions on the main stream's data format.

## 3 Data Representation and Encoding

Due to the heterogeneous nature of the data in the Teuchos platform, only part of it can be encoded in XML following the TEI P5 guidelines, most notably manuscript descriptions and transcriptions. The latter will be discussed at length later in this chapter. Especially where we handle digital images, further structures besides the means of reference foreseen in TEI are required, and collections such as e.g. the watermark albums are outside of its scope of representation. Since not all materials will be available for all manuscripts, the platform is implemented as a robust system able to accomodate incomplete sets of data.

### 3.1 Manuscripts: Digital Images

Not all descriptive materials in Teuchos are accompanied by digital images of manuscripts. In the cases where we can provide such images, each of these is ac-

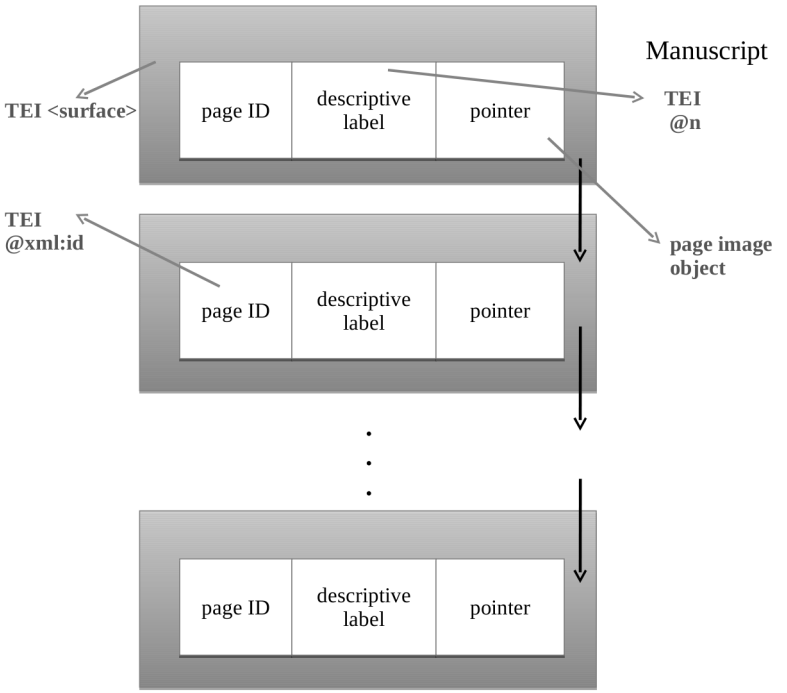


Figure 2. Linking page information within a manuscript object.

companied by a set of descriptive metadata as well as authority data. This per page image metadata will be stored in a Fedora object together with a thumbnail of the image for faster access e.g. for search processes and preview purposes.

For each manuscript of which we provide digital images, a reference document making use of the TEI <facsimile> element is created. In our case, this consists solely of a list of <surface>s, providing a unique identifier in the form of an xml:id attribute, and a label discussed below by using the n attribute. For each surface, a pointer to the Fedora object just mentioned containing a thumbnail and reference to the full image data is added.

We try to provide mostly unambiguous human readable labels for all pages (usually the main foliation or pagination and the customary short designations also used in manuscript descriptions for flyleaves etc.). Where deemed necessary (e.g. for images of the binding), a more descriptive label to be used in the display of images might be additionally supplied using the <desc> element.

To ascertain that the page structure of the manuscript at the time it was digitised is represented, <surface>s are listed according to their physical order in the manuscript, and in the case of pages not digitised, an empty surface element will be included (since we do not carry out all digitisations ourselves, there have been cases where e.g. fly-leaves have been omitted).

While the use of this minimal <facsimile> document is quite limited, it separates information on image files; their location and availability from the structural data contained within the transcriptions discussed in the next chapter, and at the same time holds the basic information required for a simple browsable presentation providing minimal navigational aids. This ensures both that the image material can be made accessible independently of the (partial) completion of a transcription (and thus prior to any in-depth analysis of the textual content), and that no images of matter that is deemed to be outside the scope of the transcription are excluded through mere omission of reference.

If images of manuscripts not available through Teuchos are made individually accessible by a third party, reference to such images could be made through this same abstraction mechanism.

**Manuscripts: Physical and Textual Structure**

To facilitate users' access to digital images of a manuscript, we also provide at least a minimal (i.e. possibly empty except for pagebreaks) transcription of the manuscript, containing structural information that can be used to offer alternate representations and improved navigation for browsing, and to give a clearer indication of the part of the text to which an image viewed pertains. Such data is encoded within a TEI <text> element. Similar information will of course also be included in full transcriptions even of manuscripts for which we cannot provide images, but we will limit ourselves to discussing merely the structural elements in this paper.

We should like to note that the structural data we discuss in this section does not necessarily presuppose the existence or public availability of images of the manuscript or manuscripts referenced. While our basic structural data can with relative ease be extracted to a standard format such as METS for export, this is only meaningful where digital images can be referred to, and information on the precise placement of structural elements within a (partial) transcription is still lost this way. In this regard, our approach is orthogonal to models that treat the page image as the main point of reference for the structural metadata.

**Page, Folio or Quire Structure**

Reference to the digital images of individual manuscript pages is made through the inclusion of <pb> elements using the corresp attribute to point to the unique identifiers established for the representation of each page (cf. all three examples below). Foliation or pagination information, however, is encoded separately using the <fw> element. This

permits recording whether numbers provided by the transcriber are actually present on the page or not, and also recording more than one such reference system (besides misfoliation and the like, the presence of a second, alternate foliation is not uncommon) through consecutively including more than one <fw> element (cf. ex. 1 and 3). There is the slight problem of clearly disambiguating multiple foliation systems. While the difference is usually evident from the placement attribute (though one may have to introduce very precise values for this attribute e.g. to clearly separate two foliations indicated one above the other), the use of this positional data for classification is less than ideal. Since alternate numbering schemes will usually be supplemented with some typographic symbol by the scholar working on the description or transcription of a manuscript to make them easily distinguishable to users of the materials, our solution is to always supply such 'normalised' forms of any numbering scheme in the n attribute. For verso pages that usually show no numbering, an empy <fw> element is used to supply the number.

Since the <facsimile> element's structure is thus kept independent from that of the transcription, the transcription can without further ado ignore the manuscript's current page order where it does not correspond to the original arrangement (e.g. through misbinding or in the case of the underlying text in palimpsests). Additional information on a manuscript's structure can also be included. Most notably, further <pb> elements may be included for (physically, not digitally) missing pages with or without loss of text (information that is not available within the limits we imposed on the <facsimile> structure above). Where a manuscript exhibits quire signatures, we also include these using <milestone> elements (with a value of "quire" for the unit attribute).

**Content-related Structure**

The most important means of providing structural information within the scope of a transcription are (1) relating the text to the structure of an existing edition of the work in question, and of course (2) encoding any structure evident in the witness being transcribed itself. Whereas in the former case, these structures should not be encoded hierarchically for evident reason (cf. below), a structure derived from the witness itself would prima facie be deemed safe for hierarchical encoding. However, this no longer necessarily holds true once transcriptions from more than one witness of the work in question are to be joined, e.g. where chapter divisions are inconsistent among manuscripts as one of the most obvious cases. To be able to retain per-witness structural information in a joined document, we therefore propose to encode all structural information using empty elements, i.e. <milestone>s. When such a joined document is edited further to become a new edition of a work in its own right, the editor(s) may (and in most cases probably will) of course decide to create a hierarchical structure taking into account the structure of the various witnesses, but this should be a later step. To avoid confusion, we should state that we do not intend to provide dynamically

generated editions. While the semi-automatic joining of transcriptions is a first step towards creating a digital critical edition, the further steps require substantial scholarly intervention.

**Content-related Structure Derived from Editions**

It is established practice to use structural information derived from an existing edition for new work on a text. While such structures may in whole or in part be of a hierarchical kind (e.g. book, poem and line number for the works of a poet), this is not always the case, be it that the final unit is the line number in a prose text (which might switch in the middle of a word where hyphenation is used), or that the entire reference is based on pages and lines of a previous edition (e.g. for most of Aristotle's work reference is to the pages, or rather columns, in Bekker's edition, followed by a line number). Most of these reference systems use two to three hierarchical levels. We propose the use of <milestone>s using a special value of "external" for the unit attribute to prevent confusion with any intrinsic references, and a special value of "canonical" for the type attribute (cf. ex. 2). The edition and numbering scheme being referred to will be indicated using the ed attribute (thus including references to more than one edition or to multiple numbering schemes derived from one edition is possible), and the hierarchical level using the subtype attribute, for which we propose using a simple range of values "level1", "level2" etc. to facilitate processing. The actual reference is supplied through the n attribute (while references are usually some kind of numerical value, extrinsic short titles or headers might in fact be encoded this way, too). Obviously information on what each value used for the ed attribute refers to and what kind of reference (book, page, stanza, line etc.) the various subtype attribute values represent for this edition need to be supplied in an <encodingDesc> so the information can be displayed in a meaningful way. The granularity of the information encoded is up to the transcriber, and might range from a mere indication of chapters, poems or similar through indication at the precision of a line number at the beginning of each page to a per-line matching between text transcribed and canonical edition. As a special case example, we have provided references to line numbers for an (otherwise unstructured) new version of a text for approximately every 5th to 10th line where this version actually (partially) corresponds to the edition of the previously known versions (cf. ex. 3).

**Content-related Structure Derived from the Witness**

The same system may be used for structure derived directly from the witness (e.g. book, poem or chapter numbers, cf. ex. 1 and 2). Due to the problems of joining transcriptions discussed above, the values suggested for the unit attribute in the TEI guidelines should again be shunned, and we propose use of the special value "internal" instead, of values "present" (for numbers actually present in the text) and "implied" (for those numbers supplemented from the evident sequence even though they are not or no longer evident in the witness) for the type attribute, again indicating the hierarchical level (where ap-

plicable) through the subtype attribute, and of course adding the appropriate reference identifiers to the <encodingDesc>. (And further noting that in spite of its name, the ed attribute can be used to indicate a manuscript siglum instead of an edition in this case. Should a manuscript contain more than one such numbering scheme, a concurrent scheme would have to be indicated through the use of a modified siglum). This intentionally does not take into account where and how the given numbers are actually indicated in the manuscript, as we expect to indicate these numbers in a normalised form. (They might be written textually rather than as numbers, or in Greek or Roman numerals, at the front or at the end of a title, in the margin or in some other location, which would all yield different representations in the transcription.) It is important to note that our approach separates the numbers from textual titles of headings (on the latter cf. below), thus allowing us to gracefully handle the case where identical titles received different numberings in different manuscripts.

**Headings and Other Non-numerical Indicators of Structure**

Where a manuscript contains chapter (or book, paragraph etc.) headings, we encode these using the <head> element (cf. ex. 1), using an appropriate type attribute to identify the kind of scheme (and to distinguish e.g. a spelled out numbering system from the written titles, if so desired), employing the same subtype and n attributes as with the corresponding numbering scheme (one might argue that this information is superfluous, as we always include a milestone for consistency anyways). Generally, a suitable numbering scheme as described above will be added in case no numbering is employed in the manuscript. In cases of a manuscript mostly following the textual structure established in an edition of the same text from other sources, but omitting the headings, these can usually be included as <supplied> within a <head> element. Should the manuscript being transcribed contain the headings in a nonstandard form (relative to the tradition of this text, or possibly to an established edition), it may be desirable to offer the standardised form, in which case the <head> would contain an <app> element with the standardised form indicated as the <lem>. (We should clarify that we use the parallel segmentation approach for all our transcribed or editoral texts, thus these elements will always be inlined, facilitating extraction of this information.) If the titles are regarded as sufficiently universal, the approach utilising <lem> might conceivably be used even for cases of multi-versioned transmission where no base text is chosen outside the scope of the headings. An alternate approach for this latter case would be to encode such titles as an external reference system (cf. above). We maintain that this latter method is always preferable when supplying typified titles (cf. ex. 2), which might be in a (modern) language other than that of the text (e.g. where the actual titles are incomplete or may be deemed misleading).

To address a case that might be seen as somewhat controversial, in one manuscript with otherwise very little structure, we have one or two phrases written in red ink on

almost every page, but few of them would be considered headings, the others being descriptive image labels, minor notes on the content and other materials. Given uncertainties in easily achieving a reliable classification in this particular case, we have as a first step added transcriptions of these phrases tagged merely with <hi>, which can thus be extracted and used as an additional 'navigational' aid for a reader trying to locate some particular part of the text in this manuscript (cf. ex. 3). Obviously proper classification of each of these items according to their content would be preferable, and arguably the approach used here will yield unsatisfactory results for the majority of similar cases. While we have included this possibility to allow users to quickly add a first transcription of highlighted parts of a text in order to make these navigable, we certainly recommend only using it where a proper classification of this material cannot be swiftly achieved, and we very much advise against using this for textual elements that would be considered actual section headings of any kind.

**Examples**

```
<pb corresp="#berolham512—165r"/>
<fw n="186*" placement="top right upper">186</fw>
<fw n="165" placement="top right lower">165</fw>
<milestone n="10" unit="internal" type="implied" subtype="level1"/>
<milestone n="1" unit="internal" type="implied" subtype="level2"/>
<milestone n="3" unit="internal" type="present" subtype="level3"/>
<head n="3" type="numbering" subtype="level3">κεφ. γ'</head>
<head n="3" type="title" subtype="level3">Περὶ τῆς ὑλικῆς αἰτίας καὶ τῶν περὶ αὐτῆς δοξῶν τῶν παλαιῶν</head>
```

Example 1. Otherwise empty transcription of one page containing numbering from two foliations, position within a three-level structure (book, title, chapter — not evident) and the current chapter's numbering and title as written on the page.

```
<pb corresp="#berolphill1538—326r"/>
<milestone n="Krampfanfall und heilige Krankheit" unit="external"
type="titletypefied"/>
<milestone n="108" ed="edoh–c" unit="external" type="canonical"/>
<milestone n="109" unit="internal" type="present"/>
<fw n="326" placement="top right">326</fw>
<milestone n="369" ed="edoh–p" unit="external" type="canonical"/>
<pb corresp="#berolphill1538—326v"/>
<fw n="326v"/>
<pb corresp="#berolphill1538—191r"/>
<milestone n="Hufleiden" unit="external" type="titletypefied"/>
<milestone n="109" ed="edoh–c" unit="external" type="canonical"/>
<milestone n="110" unit="internal" type="present"/>
<fw n="191" placement="top right">191</fw>
<milestone n="370" ed="edoh–p" unit="external" type="canonical"/>
```

Example 2. Multipage excerpt (note transposition of folios) from an otherwise empty transcription providing typefied modern chapter titles, divergent chapter count from edition and witness transcribed, as well as a reference to the edition's page count (not congruent with the witness' page structure).

```
<pb corresp="#lipsgab19—46r"/>
<fw n="σϛ" placement="top right upper">σϛ</fw>
<fw n="46" placement="top right lower">46</fw>
<lb n="1"/><hi rend="red">πῶς ἕλληνες ἀπέκλεισαν τὴν τρώϊαν ἐκείνην·</hi>
<lb n="2"/><milestone n="10516" ed="edpj" unit="external" type="canonical"/>
<lb n="3"/>
<lb n="4"/>
<lb n="5"/>
<lb n="6"/>
<lb n="7"/>
<lb n="8"/><milestone n="10528" ed="edpj" unit="external" type="canonical"/>
<lb n="9"/>
<lb n="10"/><hi rend="red">ἀρχὴ τῆς ὑποθέσεως τοῦ ὁρωμένου κόσμου:</hi>
```

Example 3. Beginning of a page of a transcription from a manuscript with two foliations, unclassified highlighted matter and indication of approximate alignment with a reference edition of a divergent version of the text.

### 3.2 Collections of Watermarks

Watermarks introduced by the paper-makers are evident in virtually all Western paper that has been used as a writing support in medieval manuscripts and documents. Since watermarks changed frequently, and produced paper was used up within a short span of years, these marks can be an important aid in dating manuscripts. To this end, numerous collections of watermarks from dated documents or manuscripts have been created, and many of these have since been digitised and made available online.

A basic discussion of watermark collections can be found in the paper of Wolf in this volume. The Bernstein project, which has undertaken to offer unified access to some of these collections and is working on a refined and more universal motif classification for use in the digital world, provides a by no means exhaustive list of existing collections on its website.

The collections of watermarks the Teuchos project is preparing to bring online contain tracings of marks from Greek manuscripts, and exhibit some specific features. Most importantly, watermark motifs occur in twin pairs due to the production process of the paper. (The marks are created through wire soldered or sewn onto the base wire structure of the molds used, and the established manual paper production process used two such molds in alternation between each sheet of paper.) Since our collections are based not on documents, but on manuscripts that frequently contain a number of consecutive sheets from the same source, they can usually offer both twin instances of a watermark motif. Given that the molds deteriorate over time, in some cases additional instances of a motif may be available where a degradation of some motif part is evident. It may also be possible to identify identical watermarks in other manuscripts.

We chose to encode our collection data in XML, and to model two different object classes in an object oriented approach (cf. Figure 3 and 4). Motif objects hold infor-

mation pertaining to a watermark motif as a whole, and they form a super class to instance objects that encode information on a single mark as represented by a single tracing. Since countermarks (smaller subsidiary marks placed at a different location of the sheet) are usually traced separately, we also treat these as individual instances of the main motif.

The motif object contains the main identification data on the motif, and lists of the pertaining instances, of similar motifs (in any collection) and of identical motifs (usually in other collections). The instance object contains all the data that will or may be different between two tracings (e.g. originating manuscript with information on dating etc., exact location in the manuscript), as well as links to the motif object, to the actual digital image of the tracing, and to the description of the manuscript the image was taken from. Since most of our collection contains data on the origin of the datings (specifically precise references to manuscript subscriptions) which are not strictly part of the watermark description, we extract this data and automatically create a minimal manuscript description to host that information. A user wishing to verify details on the origin of the dating indicated with a given watermark has easy access through the link to this manuscript description. An example of an instance object based on our XML schema is presented below:

```
<teuwmo:teuwmObj [...]>
<teuwmo:wmIdent wmIsCountermark="false">
<teuwmo:wmObjId>TEU_WMDesc_Aiglem2–21.xml</teuwmo:wmObjId>
<teuwmm:wmIdentification>
<teuwmm:wmIdnr>21</teuwmm:wmIdnr>
<teuwmm:wmCollection>Harlfinger</teuwmm:wmCollection>
<teuwmm:wmName>
<wmNameLanguage wmLang="fr">Aigle</wmNameLanguage>
<wmNameLanguage wmLang="de">Adler</wmNameLanguage>
</teuwmm:wmName>
</teuwmm:wmIdentification>
</teuwmo:wmIdent>
<teuwmo:wmManuscriptData>
<teuwmo:msName>Vatic. 1469</teuwmo:msName>
<teuwmo:msFolio>ff. 1–72</teuwmo:msFolio>
<teuwmo:msDate>1495</teuwmo:msDate>
</teuwmo:wmManuscriptData>
<teuwmo:wmLinks>
<teuwmo:pictureLink>Aigle–21m2</teuwmo:pictureLink>
<teuwmo:msDescLink>Aigle–21.xml</teuwmo:msDescLink>
<teuwmo:motifLink>Aigle.xml</teuwmo:motifLink>
</teuwmo:wmLinks>
</teuwmo:teuwmObj>
```

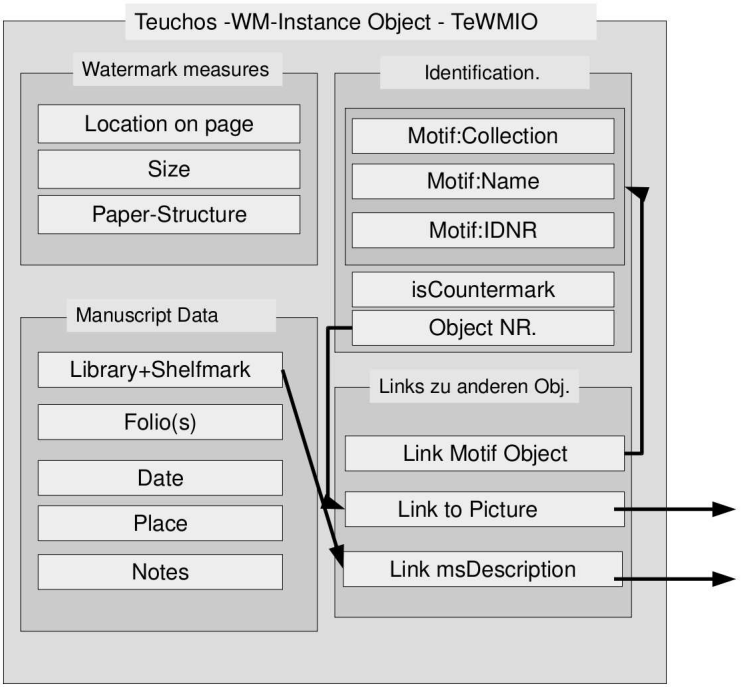


Figure 3. Watermark motif object.

## 4 Conclusions

In this paper, we introduced the Teuchos center as an evolving research environment for manuscript and textual studies. Due to constraints of space, we gave a general overview ignoring important aspects like user-interfaces, CMS, search functionality etc., and proceeded to discuss specific encoding decisions in selected areas. We chose to focus particularly on structural encoding as a precursor to preparing transcriptions, since we consider this an important step in providing fuller access to digitised manuscripts for textual scholars, and a necessary prerequisite for cumulative and shared scholarly work on the primary text sources in a distributed digital environment. As a consequence of this focus, we did not cover manuscript descriptions, for which we resort to TEI P5 as well. While the conversion of numerous existing in-depth descriptions, and the extensions of the manuscript description provisions of TEI to support all the particular aspects of the *Aristoteles Graecus* cataloguing model, as well as providing for more in-

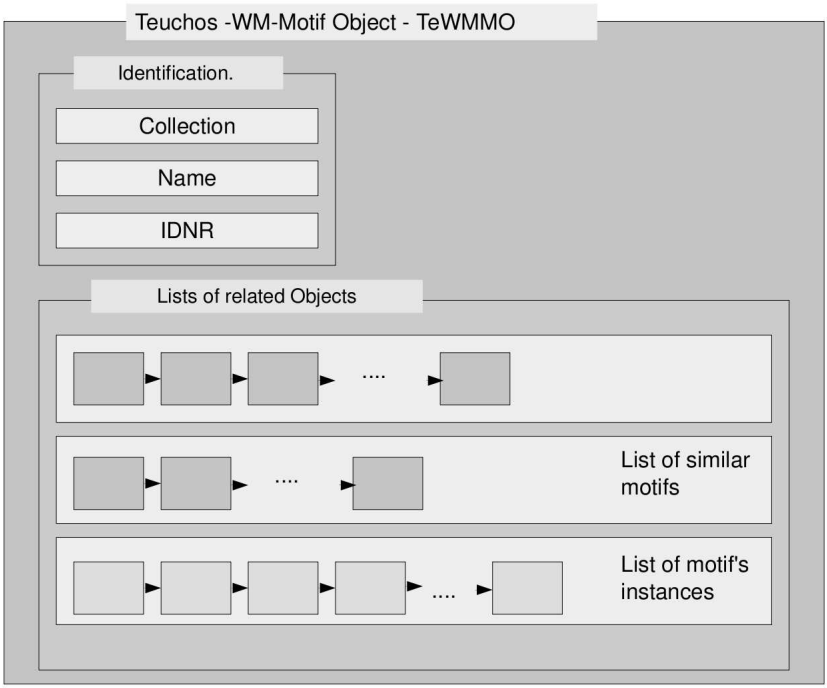


Figure 4. Watermark instance object.

depth analytical markup of the existing and of new descriptions, would be a subject for much discussion, it would have gone far beyond the scope of this paper.

## Bibliography

*Aristotelis Opera ex rec. Immanuelis Bekkeri.* Ed. Academia Regia Borussica. 4 vols. Berlin, Reimer: 1831-1836.

Bernstein project's overview of watermark collections. <http://www.bernstein.oeaw.ac.at/twiki/bin/view/Main/PaperDatabases>.

DFG, Thematic information networks. <http://www.dfg.de/forschungsfoerderung/wissenschaftliche_infrastruktur/lis/projektfoerderung/foerderziele/informationsnetze.html>.

Dublin Core. <http://dublincore.org>.

Fedora. <http://www.fedora.info>.
Moraux, Paul et al. *Aristoteles Graecus. Die Griechischen Manuskripte des Aristoteles.* Berlin: De Gruyter, 1976 ff.
Oder, Eugen and Karl Hoppe, eds. *Corpus hippiatricorum Graecorum*, 2 vols., Leipzig: Teubner, 1924/1927.
Papathomopoulos, Manolis and Elizabeth Mary Jeffreys. *Ο πόλεμος της Τρωάδος.* Athens: Μορφωτικό Ίδρυμα Εθνικής Τραπέζης, 1996.
Stevenson, Allan Henry. "Watermarks are Twins." *Studies in Bibliography* 4 (1951-52): 57-91.
Teuchos. <http://www.teuchos.uni-hamburg.de>.

## Examples

Example 1. Berlin Hamilton 512, f. 165r. The part of the manuscript in the example contains a philosophical text book by 13th century scholar Georgios Pachymeres. The two foliations hail from before and after the loss of the first 21 folios of the codex.

Example 2. Berlin Phillips 1538, f. 326r, 326v and 191r. A richly decorated 10th century manuscript containing a collection of texts on horse medicine. The chapters in the example deal with convulsions and epilepsy and with hoof diseases respectivly, reference is made to the edition of Oder and Hoppe.

Example 3. Leipzig Gabelenz 19, f. 46r. A version of the Greek adaptation of the medieval epic poem War of Troy. The manuscript contains an alternate foliation using Greek numerals. The first caption transcribed is descriptive, the second a classification of what is to follow. Reference is made to the verse numbers in the modern edition of the main tradition by Papathomopoulos and Jeffreys.