

Grant proposal

Subject:

Development of standard tools using EAD, EAC and METS

for different kinds of archival material and for different aggregation levels

on the basis of the MEX tool set prototype developed by the <daofind> project Nov 2004 to June 2006

Summary.....	1
Aims.....	2
- Holdings guides.....	3
- Finding aids for case files - standard version.....	3
- Finding aids for case files - extended version.....	3
- Finding aids for collections of personal papers.....	4
- Finding aids for collections of personal or persons related files.....	4
- Viewer for digitized records.....	4
The Software Basis for the Tools Development.....	5
The METS Viewer.....	9
Results.....	11

Summary

With support of The Andrew W. Mellon Foundation the project <daofind> was able to develop the MEX tool set as a prototype for a new kind of archival working tools, and a pilot application of a presentation model for textual archives with the presentation of digitized archives integrated in online finding aids on item level and on collection level using the international standards EAD, EAC and METS. Both demonstrated the feasibility of the approach. MEX shall now be used as a development platform for specialized tools.

The aims of the new project consist of developing dedicated tools for special tasks on the basis of the MEX tool set prototype in order to support and enhance the

production of critical masses of digitized archives in the Internet with little investments in resources, especially personal work force for repeated tasks. The finding aids prepared and edited with those tools and enriched with digital reproductions will be ready for up load on a website, for integration into search engines like MidosaSEARCH used by the Federal Archives for simultaneous cross searching and structural navigation and for integration into other central access points. The tools are delivered with built in style sheets for standardized presentation models and a HTML preview function. As the users work with and produce normal XML files the results are open for other presentation formats and other style sheets may be chosen for their presentation.

Aims

The MEX tool set prototype is based on the three standards EAD, EAC and METS and produces valid XML documents. They offer the whole flexibility of the standards and are powerful and highly complex tools. Therefore they offer a great flexibility and variety of possible implementations, which causes on the other hand the need to choose. The new project aims at reducing the complexity by building a set of specialized tools for typical working scenarios. Part of the complexity is eliminated by making important choices before starting the description. If EAD can be used for different forms of finding aids concerning different material or different aggregation levels, normally these are not changed in one EAD document and therefore this approach can be selected before starting the description. Also the choice whether a person, a corporate body or a family should be described as a records creator with an EAC document can be done before starting to capture the data. The differences between the standard tools will result from different subsets of the corresponding standards, from a specialized labeling on the work screen, which can be switched to different languages, and from adapted transformation style sheets for HTML presentations.

The standardized tools will support the integration of links between different XML-files needed for a combined presentation e.g. of EAD with EAC for holdings guides and of EAD with METS for the image viewer integrated into finding aids.

The following six specialized tools will be produced during the project:

- for holdings guides on collection level,
- for finding aids for case files on item level
 - in a standardized and
 - in an extended version,
- for special finding aids describing collections of personal papers,
- for finding aids for personal files, and
- for a viewer of digitized images integrated into finding aids.

Default style sheets for presentations similar to those used already for online finding aids created with MidosaXML will accompany the tools for the establishment of different types of finding aids. The images viewer will consist of a style sheet creating the specialized presentation format similar to the one of the pilot application of the daofind project together with an editing tool for entering all data needed to define the

internal structure of each digital archival object which corresponds to one descriptive unit.

- Holdings guides

Holdings guides contain short and comprehensive overviews on all collections of an archival repository including their internal relationship in a sort of classification scheme, the so-called tectonics. They are built with a combination of EAD and EAC. EAD will be used for describing the hierarchy of the structure together with information on the collections themselves. EAC in the three versions for corporate bodies, for families, and for persons will be offered for information on the records creators with adapted subsets of elements each. EAC offers data elements usable for an authority control of the official name and other names used and allows describing the competences including their development over time. In larger organizations the EAC files might be maintained by decentralized specialists, while the EAD backbone is maintained by a central service. This tool will use subsets of EAD and EAC, an adapted labeling, and a style sheet which will create the combination of the EAD and EAC files necessary for the presentation of this kind of document.

- Finding aids for case files - standard version

The description elements for standard finding aids are reduced to the main parts of data on the descriptive units, while in the introduction or in information about the processing (e.g., the <processinfo> on the highest level of the structure) general information for the whole collection might be given. This tool will offer only the core elements of finding aids and therefore it would be adapted for implementation also at workplaces for less qualified personnel. The resulting XML documents can be opened with the extended version, if the need should occur to add more differentiated information at any place inside the document. This version uses a reduced set of elements of EAD and the normal labeling. No special style sheet is needed.

- Finding aids for case files - extended version

The extended version offers more possibilities for the description of the single items, for supplementary information, and for documentation on the processing as well as on appraisal decisions on different levels of the hierarchical structure of the finding aid. In contrast to the standard edition more choices are offered so that professional and experienced personal can use more sophisticated descriptive methods. This version uses a larger set of elements of EAD with the standard labeling and the standard style sheet like the one for the MidosaXML online finding aids, which offers already e.g. places for the documentation of the processing or the appraisal decisions.

- Finding aids for collections of personal papers

The description of collections of personal papers often needs more place for biographical information on the person and on the activities during which the papers

of the collection were created and assembled than standard finding aids offer. Therefore a different subset of EAD tags is needed. Even more relevant is the labeling for the different parts of the finding aid in the tool, so that the describing archivist finds adequate places for all the knowledge about the person as well as the creation and growth of the papers. This version will use an adapted subset of EAD for describing the collection itself and of EAC for the records creator. The combination of EAD with EAC for persons mentioned in the collection will be integrated in such a form, that the EAC data can be extracted separately, compared to and used to complete existing authority files. A different labeling and an adapted style sheet with a combination of EAD and EAC is required.

- Finding aids for collections of personal or persons related files

Personal records are created by activities concerning the administrative needs of a certain person or a group of persons. The files are normally arranged according to the names and other identifying data like the date of the birth in the form of a series. Such files can be described in a very standardized form. In addition to the title of the series, the description needs the identifying indications and often more information from the files is wanted for the finding aid. Here also a reduced subset of EAD can be applied. EAC will be linked to the EAD data and will open the possibility to use an interface to authority files. This tool again needs a separate set of EAD elements as well as an adapted set of EAC elements with an own labeling and a specialized style sheet with more information categories, e.g., biographical information on persons.

- Viewer for digitized records

Digitized images will be linked to the finding aids with the help of a viewer based on the METS file containing all their metadata. The METS file combines all images from one descriptive unit to one digital archival object with the possibility of editing their internal interrelations. A METS file will be created automatically by reading all the addresses of the images with the help of a wizard. The corresponding style sheet generates a presentation which is linked to the descriptive unit in the corresponding finding aid so that it opens on a click and shows the images one after the other for browsing. The structuring capacity of METS can be used, if wanted, to point the visitor to the most relevant pages first and give more orientation for navigation through the digitized files.

The presentation model for online finding aids, which was developed together with the tool MidosaXML will be the default solution delivered with the tools. The descriptions of the units will offer the possibility to link to METS files for images. The online finding aid presents a set of three frames. A left frame contains the expandable structure, like a table of content in a book with links to the text of the document, shown in the right hand main frame on a click on the respective header. The group of descriptive units called up this way is presented as a whole in the scrollable main frame like a chapter in an e-book. Above a frame shows the complete hierarchy going up to the top and starting with the selected group. This frame has a function similar to page headers in a printed book. All three frames are synchronized. That means that also the page header and the navigation frame are updated with changes of the content in the central frame and vice versa. This presentation model

allows using and combining four different search strategies into one search history. The four combined strategies are a structural navigation, a full text search, a browsing search through the whole finding aid and the use of index terms, which leads directly to the place named in the reference.

This standard model for online finding aids was taken as the bases for the integration of images. If a METS-file with metadata about images is available and its address is indicated in the <dao> element of the EAD file, during the generation of the HTML files a link is created with a text entered in the dao description of EAD file, which in the presentation opens a new window with the viewer for the images.

The viewer with the pilot presentation model is demonstrated on the project web site of the daofind project (www.daofind.de) with three exemplary finding aids enriched with 30.000 images from the three record groups. Two complete record groups and parts of a third one were chosen as a sort of test bed for the pilot application. The first two record groups are those of two secretariats in the central committee of the former socialist party (SED) of East Germany (DDR) with material from 1945/46 until the mid 50s which is historically an extremely interesting period before the actual creation of the East German state. The third one is the fonds of the central party control commission of the SED, a disciplinary court inside the same party, with about ten files on general political matters. All images were digitized from microfiches, which had been produced earlier for preservation purposes. The finding aids are encoded in EAD. Each descriptive unit is linked to the corresponding METS file with the button "Akte einsehen" ("inspect the file").

The Software Basis for the Tools Development

The tools development will start from the MEX prototype of the daofind project. The software is conceived as a plug-in to the open source Java development platform Eclipse. It is based itself on Java and therefore it is platform independent. It runs on different operating systems like Windows as well as Linux. It can also be used on Mac OS X with some little restrictions. A full compliance to the Apple operation system depends on further developments of the Eclipse platform, which can be expected in the near future.

Eclipse provides a very innovative environment, which had been developed by IBM, who opened the source code for the open source community. It evolves still rather fast, and has a high potential for the future development of Internet technologies. However it is used in an increasingly broad way so that it is stable enough to serve as the basis for professional working tools.

The prototype for more specific tools integrates certain functions that are usually not available for XML editors like an integrated HTML view and a function for the fast generation of the sets of HTML files needed for the Internet presentation of an XML-document. These functionalities will be handed over to the dedicated tools that will be built during the new project.

The prototype MEX is installed as a plug-in to Eclipse. When starting to create an XML file corresponding to one of the standards the corresponding schema is copied into the workspace. The editor opens an XML document with just the set of elements needed for validation. The other elements can be added as they are wanted. All three

editors in the MEX tool have been developed with the guiding principle, that they would not require knowledge of XML for their use but should be instead based on common professional terms and concepts and that they should be adaptable to different technical environments.

The aim is that they can be integrated into the normal daily work at the desk of the archivists allowing them to do their work in a new, more output oriented way. No special staff should be needed to prepare the Internet presentations from the archival working products. Instead integrated work processes should be made possible, producing Internet presentations as well as offline publications like CD or DVD and printed finding aids with them same data that is captured once and needs no manual transformation or reworking for different publication formats.

The tool set MEX works on document level that means on an aggregated level above the technical XML level. The implementers of the tool construct documents like finding aids according to their professional competencies and use the language they are accustomed to. Under the visible surface the terms from the professional archival language are automatically translated into the corresponding XML schema. A 1 to 1 relation between the natural language terms and the XML elements is not needed. The terms can be translated into single tags, into combinations of tags or into combinations of tags and attributes with special values. This flexibility offers the chance to easily change the terms used on the working screen, if they have the same function but perhaps another wording e.g. for special kinds of finding aids for certain archival material, like collections of personal papers in comparison to files from a state agency. In the same way the language can be changed to created versions for different nationalities. The specialized tools that will be developed with the new project will contain the different professional wordings needed for the different applications.

The information about the use of the tags and of their labels on the screen are captured and documented in a proper XML file that contains all the definitions needed for the actual application of any element or attribute of the corresponding schema. The definition files for the translations from the professional language to the language of the XML standard, offer at the same time the possibility to define the restricted subsets of the standards that are applicable for special kinds of finding aids. Certain elements and attributes with their values can be left out, if they are not needed. Then they are not presented on the working screen, and do not seem to request to make choices that are not necessary for this special form of finding aid.

The XML-files with the definitions are used to build up the working screen. Therefore the working screen presents just that subset of the standard that is needed to work with in this working situation. However the describing archivist is still free to use just those elements of the finding aid estimated as useful for this special case of descriptive work. Each document created with that tool is validated against the complete corresponding schema when it is saved.

As Eclipse offers the possibility to use different views on a screen, the working screen with the editing surface in the middle frame shows an editor view. It contains the fields that are filled with content in a structured manner. In a tool box view on the right hand side sections and elements of the structure of the respective professional documents, e.g. finding aids, are offered and a click on a term enters the

corresponding field into the middle frame, the actual work area. Here the content is entered or edited. On the left hand side in an outlook view the structure can be followed as it grows or is changed. Other views show results of validation or help texts corresponding to the selected elements. All views can deliberately be opened or closed. The views of the document in the middle frame can be switched to HTML or XML so that they allow following the actions and their results closely.

The tools include an import functionality that reads every document, which is valid and conforms to the corresponding standard. Those elements and attributes that are not defined in the XML definition files will be indicated with their official tag name in brackets, so it is clear that they cannot be edited. Nevertheless the XML file can be transformed with the integrated style sheets e.g. into the default or another HTML presentation, for which a style sheet is integrated. New elements can be added like links to images or other XML files and the files can be stored as valid XML file with the added fields. The style sheets will be written in XSLT 2.0 and allow generating a HTML presentation and the production of a text file as a manuscript for printing.

The tools that will be created during the new project will be ready for use in typical working situations when describing archives. They will contain the editing tool, which works like an XML editor but offers professionally named elements of the documents to be produced. They include the import functionality, which is open for any valid document conforming to the respective standard. And they will contain default style sheets for online finding aids and printing. The lay out of the online finding aids will differ from each other a bit corresponding to the special forms of the finding aids to be produced. So the specialized tools will be delivered with slightly different style sheets. These tools will be distributed freely and the source code will be laid open.

The flexibility of defining the needed subset as well as the terms and language of the surface was developed during the first project and was made possible with a supplementary funding of the costs by the Bundesarchiv itself. The software developer has found out innovative programming techniques for this part and he has invested supplementary resources of his own with the intention to acquire innovative competencies for his own business. Therefore we suggest that the function needed for the definition of the subsets of the standards and the labeling remains licensed, while the rest, that means the complete programming code of the specialized tools will be opened. The freely delivered tools include the unrestricted import functionality as well as the default style sheets and the openness for using other style sheets.

The new tools will be independent and autonomously applicable instruments for the creation of XML files conforming to the standards EAD, EAC and METS. The XML-files can be changed and altered with any other editor. The tags needed are offered and can be chosen as wanted. The selection and the display of the tags for the different applications is defined using the special definition tools of the prototype, which remain proprietary and which are not needed for the working tools. The tools will be designed in such a way that the tag table can be created and changed manually.

The tools can be used outside the Bundesarchiv's prototype on any platform (Windows, Linux or Apple Macintosh) of any computer, as they will be written in Java and are based on the Java development platform Eclipse. They can be applied as stand alone systems or can be integrated into networks and use common storage

space. They do not need database management systems in the background and instead use files systems for the storage of the data. The products in the form of automatically generated HTML files can be integrated into websites without further adaptation. So they need little to no IT personal resources for their application and for the presentation of finding aids and images on the Internet and for the management and maintenance of the data.

The EAD editing functionality of MEX allows using the functionality of the <c>-elements for the creation or representation of hierarchical structures conforming to the definitions in the EAD DTD. The structure can be changed by moving any heading or descriptive unit to another place. The EAC editor can be used to capture all information on records creators including different forms of its names and the times, when they were used. The records creator can be persons, families or corporate bodies. Competencies can be listed and described together with the time period during which they were existent. Records creators can be corporate bodies, persons or families. Therefore this standard is well adapted for parts of the information delivered with the standard form of a holdings guide, were the record groups and their provenance are described for a general overview of the complete holdings of a repository.

For the tool for creating holdings guides one EAC file will be used for the description of each records creator, and it will be linked to the EAD file used for the structure of the holdings guide demonstrating the tectonics. The linking could be done on the fly, so that the maintenance of the content can be decentralized. This feature will be developed by the project. How far the user should be aware of using different standards is not yet decided. Easiness of use and necessary orientation shall be respected while conceptualizing this tool.

The editing function for METS files can be used to create semi-automatically valid METS files, which can directly be linked to EAD finding aids. All addresses of files with digitized images that should be grouped together inside the METS-document can be entered automatically into one file group section of the METS document by selecting the corresponding folder.

At the same time one or more structural maps are created, which is another required part of a valid METS-file. Having done this the file would be ready to be linked to an EAD finding aid. All digital data of one descriptive unit are combined to one digital archival object and can be presented with a link form the corresponding online finding aid. If wanted more metadata may be supplemented manually with further indications that control the choice of an adapted style sheet or of certain behaviors of the images.

The underlying concept supposes that one METS document corresponds to one descriptive unit in an EAD file and combines all digital files, if useful in several format groups e.g. for different dissolutions or including OCR transcripts together with the images.

The METS Viewer

As we have not found a special viewer for textual records the daofind project tried out to realize a presentation model especially adapted to archival records. This

specialized viewer for METS files with the metadata of digital archival objects that correspond to one descriptive unit on the lower level of the EAD hierarchy tries to allow a fast estimation of the need to look at a page in more detail and to give better orientation while browsing through large files that may contain up to 500 pages. The archival objects represented by a METS file consist of a certain number of documents related to each other like a chain of communication events, containing incoming letters and answers to them or reports and reactions, drafts and the finalized copies concerning the same case. METS is often used in the Internet for the presentation of pictures or documents of a single or some few pages, like handwritten letters, postcards or similar pieces. Normally the images are first shown as thumbnails that can be opened in a bigger size by clicking on them. This presentation is not the best one for a collection of 100 or 200 or even more images from the same file with little external differences. The intention when thinking about a new presentation format was to show a reduced reproduction, which nevertheless gives enough information to know whether it is worthwhile to click on it for opening it in a larger size. The external shape of the documents, which can be identified on a small image does not give enough information for this decision. Yet the upper part of a document normally contains information about the sender, about the dates and the subject and shows the beginning of the text. So it seemed to be more helpful to have a glance at just this part of a document to estimate its relevance for the own information needs.

The second reflection was, that for many multipage documents in a descriptive unit, like minutes or reports, it is sufficient to see the first page to decide whether the whole document should be looked at. So a selection of these first pages and other structural relevant pages from the folder, e.g. those initiating a new action or a new chain of communication helps the reader with identifying the relevance of the whole report or minute for his research.

Therefore the METS viewer contains two layers. The first one is conceived as an orientation layer containing the upper thirds of those pages that are selected as useful for the readers for a first estimation of the relevance of the papers for their research. These upper thirds of the pages are shown in scrollable windows on one page and a click on them opens the full size image on the browsing layer, where a second click opens the next page. Numbers of the pages on top as well as arrows support quick browsing. A link to the orientation layer brings the reader back to the point where the browsing started. This presentation is directly generated from METS without intermediate transformation of the data.

The selection of pages for the orientation layer is done by the describing archivist after the creation of the present METS file with the automated integration of all addresses of all digital files and the creation of the corresponding structural map. It is done by manually entering certain attribute values for the second <div> element in the structural map of the METS file. On the working screen the archivist uses the label "Type of record" and selects one of 8 types, e.g. "incoming letter" adding to it a header for display under the image describing this document. In the XML document the following combination of the elements and attributes is created for this term on the working screen following the integrated definition set on the basis of the METS schema: <structMap><div><div>TYPE="incomingletter". The first upper <div>

element under structural map shall be usable for the selection of other presentation models for different kinds of descriptive units.

The concept for this working tool as for the presentation model is completely new. It encountered much support from the public when we showed it during presentations. The presentation model can be tried out on the webpage of the daofind project with the 30,000 images in the three pilot finding aids. The presentation pages will be completed with more context information on the corresponding descriptive unit. On the website there is also an English slide show available that shows the working screen. With this implementation of METS the metadata can directly be used for the HTML presentation, which can be controlled during the editing of the METS file. A profile for this specific use of METS will be defined and submitted to the METS editorial board.

All tools that will be developed with the new project present special functionalities that go far beyond the functions of simple XML editors. The main feature is that the work is done on the level of the document and not of the XML schema. No tag names have to be learnt and they need not to be shown on the working screen. The labels used are terms coming from the professional language and they are translated into the corresponding combination of tags and attributes. When starting the tool with one of the standards, all elements of the document needed for a valid XML file are entered. Elements can be moved to another place with drag and drop as far as the schema allows it.

While different views allow switching between the editor view, an HTML and the XML view, changes can only be made in the editor view to assure consistency. With buttons on top of the working screen the hierarchy of the documents can be deliberately expanded or collapsed.

The METS editor is completed by a picture view that shows the corresponding image. It is synchronized with the editor of the XML file. The picture view can be expanded to the size of the screen for browsing through the whole digital archival object and to stop when the need is felt to add further metadata, especially for structuring it and controlling the presentation on an orientation layer.

The technological basis of the software will be adapted to the new developments with Eclipse 3.2 and the newly released standardized version of XSLT 2.0. The links between the XML documents, needed for building up the network of HTML pages for the presentation will use the XLink name space.

The working documents are stored and used in XML format. There is no translation needed to a relational database format. So the tools are very fast. In contrast to XML editors an HTML export is integrated as well as an export into a word processing format. The HTML export permits to produce at once a presentation for the Internet that can be uploaded to a server without changes. All documents can be linked to the other documents and be used for the presentation of a network including the holdings guide with linked finding aids offering access to the images from the record group.

One further step will be the expansion of the presentation model for digitized archives by a full text search in the images. Therefore OCR or copied texts will be used for searching and the results will call up the corresponding images. The text will just be used for searching and will not be shown, so that a higher degree of errors is

tolerable, especially considering the fact that in any case the full text search can never replace completely the structured navigation and the use of interrelations for research in archival material.

During the time of the daofind project the federal archives have installed a new search engine for online finding aids, MiodsaSEARCH. It is based on Lucene and allows to search across all integrated finding aids. It combines the full text search in the finding aids with a structured navigation on both levels of the holdings guide and of the single finding aids. So the single finding aids can be found either with the full text search or by navigating through the tectonics. A group of finding aids can be selected for joint searches and the results lead directly to the place inside the finding aids or the holding guide where the chosen term is used. This search engine integrates finding aids coded according to EAD. For the moment 200,000 descriptive units are searchable. Early in 2007 it will be the main access point for all online finding aids of the Federal Archives with about 1 mill descriptive units in 6000 record groups. This search engine is delivered for free by the Archives School and Institute for Archival Sciences in Marburg, the central training institution for archivists in Germany as a light version. The server based version of the search engine for larger amounts of descriptive units is licensed by the developer. The Archives School Marburg is also willing to assure together with the federal archives the distribution of the new tools and will use them in the training courses as well in the seminars for continued training.

The Archivist's Toolkit project seems to offer some functionalities for the managing of access terms and authority control especially with interfaces to databases of authority terms that might be interesting additions to the tools. As there is little experience with this type of descriptive work in German archives, but the need is felt to be more attentive to these approaches, we would like to examine more closely the possibility to re-use and take over certain functionalities from the Archivists Toolkit project into the new tools. This seems even more interesting as the Archivist's Toolkit is compliant to EAC in this part. It will be necessary to analyze where possible interfaces may be. The main difference between both approaches is that MEX works directly with the internal connections of XML without translation to relational database structures and therefore has less performance restrictions.

Results

The results of the project will be a set of descriptive tools for special archival processing scenarios. The describing archivists do not need to learn the tagging of the standards but can work in their usual professional environment. The language used on the working screen applies the professional terms and instead of XML-elements the user works with elements of the documents that are going to be produced. These tools integrate the production of Internet presentations into the daily archival work and therefore they can help to produce larger quantities of online-finding aids with greater masses of integrated digital images in less time than today. No special IT staff is needed for the preparation of archival descriptions for Internet presentations, because it is part of the normal working processes. The tools will be able to reduce considerably the main part of the costs for digitizing archives, which are constituted by the personal resources needed for description and capturing of

metadata. The specialized tools will produce valid XML files conforming to the requirements of the three standards. They are open for presentation with different style sheets for the HTML web presentations of the repositories or for transformation to share the data with central access points and to integrate the data into databases. The implementers are not dependent on proprietary software for a database as all data can be maintained in files systems or they can be managed with XML databases that control the work with the files. The tools shall be offered for download as open source software, so that they can be adapted to individual needs and style sheets can be added, altered or exchanged.

The project will need about one year so that it will be completed in October 2007.

By April 2007 first results should be delivered on CD to the participants of the 3rd European Conference on EAD, EAC and METS, which is organized by the Federal Archives in Berlin during the German presidency of the European Union from April 24 to April 26, 2007 under the title "International Standards for Digital Archives" (cf. www.instada.eu). The project will be presented and explained in a contribution to the conference and with a booth.

Prof. Dr. Hartmut Weber
Präsident des Bundesarchivs

Berlin, Oct. 3, 2006