Submitted on: August 13, 2013

The ALTO Editorial Board: Collaboration and Cooperation across Borders

Frederick Zarndt
IFLA Newspapers Section, Coronado CA USA
frederick@frederickzarndt.com

Joachim Bauer
Content Conversion Specialists, Hamburg Germany
j.bauer@content-conversion.com

Markus Enders
British Library, London UK
markus.enders@bl.uk

Brian Geiger
University of California Riverside, Riverside CA USA
bgeiger@ucr.edu

Kia Siang Hock
Singapore National Library Board, Singapore
siang_hock_kia@nlb.gov.sg

Jukka Kervinen
National Library of Finland, Mikkeli, Finland
jukka.kervinen@helsinki.fi

Evelien Ket
Koninklijke Bibliotheek, den Haag, the Netherlands
evelien.ket@kb.nl

Jean-Philippe Moreux
Bibliothèque nationale de France, Paris France
jean-philippe.moreux@bnf.fr

Nate Trail
Library of Congress, Washington DC USA
ntra@loc.gov

Abstract

Many library digital text collections created from pre-digital era materials (books, journals, magazines, etc) and nearly all library digital historical newspaper collections use digital mages (TIFF, JPEG, JPEG2000, etc), OCR software (ABBYY FineReader, Nuance Omnipage), METS XML, ALTO XML, and various metadata standards (MODS, Dublin Core, PRISM, MIX, etc). Both METS (Metadata Encoding and Transmission Standard) and ALTO (Analyzed Layout and Text Object) are XML standards developed by the international library community and administered (hosted) by the Library of Congress at http://www.loc.gov/standards/mets/ and http://www.loc.gov/standards/alto/ respectively.

The current editorial board has members from the National Library of Finland, the British Library, Singapore National Library Board, Bibliothèque nationale de France, Netherlands Koninklijke Bibliotheek, the Library of Congress, the University of Kentucky, the University of California Riverside, and a software company, Content Conversion Specialists. All but two are IFLA members, and several serve on other standards boards in addition to the ALTO board. (You can see the list of current editorial board members at http://www.loc.gov/standards/alto/community/editorialboard.html.)

With members in cities that span 16 time zones, you can imagine collaboration, cooperation, and good communication are essential to achieving anything. Of course a willingness of the members in the outlying time zones to get up early or stay up late is indispensable too. Good telecommunications infrastructure is imperative, and, as we will see, free and easy (Skype) sometimes is not reliable.

This paper gives an account of the history of the ALTO XML standard, of the ALTO editorial board, and of the ways that the board organizes itself and conducts its business. The paper describes the collaborative process used by the board to receive, review, and adopt changes to the standard, and it gives especial attention to step-by-step process collaboratively developed to track and implement changes to ALTO. And last, but far from least, it informally examines members' motivations for participation in the board.

Keywords: library standards, XML standards, ALTO XML, OCR, digitization

---

1. Overview

Analysed **L**ayout and **T**ext **O**bject (ALTO) standard is a XML schema of metadata for describing the layout and content of physical text resources such as pages of a book or a newspaper. ALTO accurately captures technical details of text pages such as the position of characters, words, paragraphs, illustations, footnotes, etc. These details make it possible for access systems like Chronicling America, Papers Past, Trove, and NewspapersSG (and many others) to precisely locate and show a character, word, paragraph, illustration, or footnote on a page image.

ALTO is an open standard and may be freely used by anyone.

The Metadata Encoding and Transmission Standard (METS) is also a library XML standard often used in conjunction with ALTO. While METS XML[1] can represent the structure of a variety of digital objects (text, video, audio), it cannot, nor is it intended to, describe text components (paragraphs, words, characters, illustrations, etc) of a text digital object such as a book, magazine, or newspaper. The ALTO XML standard has been developed for this purpose. And while ALTO XML files are primarily intended for use with METS XML files, they may also be used independently.

The remainder of this paper will recount the history of the ALTO schema, the history of ALTO editorial board, the ALTO administration and maintenance process, and the operation of the ALTO editorial board. For those who are interested, ALTO technical details may be found at http://www.loc.gov/standards/alto/.

2. A Brief History of ALTO XML

METAe was a 3-year EU-funded research and development project which began in September 2000. The project was a collaboration of 14 partners from 7 European countries and the USA and coordinated by the University of Innsbruck[2]. The ALTO XML standard is one of the products of the METAe project.

One of the METAe partners, Content Conversion Specialists (CCS), administered and maintained the ALTO standard until August 2009. Then responsibility for its administration and maintentance was transferred to the Library of Congress and the ALTO Editorial Board. The version of the ALTO schema was changed to 2.0 as part of the transfer and is identical to the previous version (1.4) except for the addition of a *loc.gov* namespace URI and an updated URI import reference for the inclusion of xLink functionality.

One of the first uses of ALTO for a mass digitization project began in 2004 during the inital phase of the Library of Congress's National Digital Newspaper Program(NDNP)[3]. ALTO is an essential component of NDNP because it facilitates the capture of text reading order and word position on each OCRed newspaper page. ALTO makes it possible for NDNP access system, Chronicling America (http://chroniclingamerica.loc.gov), to show (highlight) a search term position on a newspaper page. Without this capability users would find it

[1] METS has been and continues to be developed by the METS Editorial Board whose members are drawn from the international library community. The Library of Congress administers and maintains the METS standard, schema, and documentation. See http://www.loc.gov/standards/mets/.

[2] METAe project members were Leopold-Franzens-Universität, Institut für Angewandte Informatik Universität Linz, Mitcom Neue Medien GmbH, CCS Content Conversion Specialists GmbH, Universidad de Alicante, Friedrich-Ebert-Stiftung, Cornell University Library Department of Preservation and Conservation, Bibliothèque nationale de France, The National Library of Norway, Biblioteca Statale A. Baldini, Dipartimento di Sistemi e Informatica University of Florence, Universitätsbibliothek Karl-Franzens-Universität, Scuola Normale Superiore Centro di Ricerche Informatiche per i Beni Culturali, and Higher Education Digitisation Service HEDS. Also see http://meta-e.aib.uni-linz.ac.at/.

[3] A detailed description of the National Digital Newspaper Program (NDNP) and the technical requirements are found at http://www.loc.gov/ndnp/. Chronicling America (http://chroniclingamerica.loc.gov/) is the access software system for NDNP data.

difficult indeed to locate the search term because newspapers may have 8,000 to 15,000 words on a page.

Chronicling America is one of the first and most prominent users of ALTO XML, but today it is far from the only one. There are many, many text digitization projects around the world that use ALTO XML.

3. An Even Briefer History of the ALTO Editorial Board

The ALTO Editorial Board was formed at the same time administration and maintenance of ALTO was transferred to the Library of Congress. In August 2009 it met for the first time at the National Library of Sweden in conjunction with an IFLA / National Library of Sweden sponsored newspapers symposium "*The present becomes the past*"[4]. Initially the board was comprised of volunteer library professionals with an interest in text digitization, the library standards liaison from the Library of Congress, and employees of CCS.

The authors of this paper are the current board members. As one can see, board members are a diverse and cosmopolitan bunch with members from Europe, North America, and Singapore. Board member locations span 16 time zones from UTC – 8 hours to UTC + 8 hours. Consquently, and some board members would say unfortunately, some members must get up early for a meeting while others must stay up late.

Since August 2009 the intention of the board has been to meet monthly by teleconference or web conference. In practice the board meets when a quorum of its volunteer members has free time from their other responsibilities. Sometimes the meetings are held every other month (or even less frequently). Very rarely there may be two meetings in a single month.

4. How the Board Works

> *The purpose of the ALTO Editorial Board is to maintain editorial control of ALTO, its XML schema, and official ALTO documentation. Additionally, the Board promotes the use of the standard and endorses best practices in the use of ALTO as the practices emerge. The ALTO Editorial Board is representative of important communities of interest for ALTO.*[5]

The current board has both library and industry members. All board members from libraries are consumers of ALTO; some library members are also ALTO XML file producers. Industry members are produceers of ALTO XML (service bureaus) or creators of ALTO production software (CCS). It's important that both the producer and consumer viewpoint is represented because each has a very different perspective, for example, as ALTO consumers, libraries are far more concerned about backward compatibility.

---

[4] The conference program is found at http://www.kb.se/english/about/news/Present-past/ (accessed July 2013). The conference was co-sponsored by the IFLA Preservation & Conservation Section, and IFLA Core Activity on Preservation & Conservation (PAC)

[5] If the statement of purpose sounds suspiciously like the METS statement of purpose, that's because ALTO's statement is patterned after the METS statement (cf. http://www.loc.gov/standards/mets/mets-board).

So far board members are recruited on an *ad hoc* basis, to replace members who have left the board, or to fill other needs. In other words, there is no formal recruitment policy. Recruitment policy is likely to become more formal as the ALTO board matures. Draft board membership criteria, modeled after the METS Editorial Board membership criteria, are listed in Appendix 1.

There are no restrictions on who can submit a proposal to change or extend ALTO. Since the Library of Congress began administration and maintenance of ALTO, no changes have been made to ALTO, but there are several proposals before the board. These proposals have been submitted by ALTO board members, by the IMPACT project[6], and by Bibliotheque nationale de France.

The ALTO board developed an almost self-explanatory proposal submission template (cf. Appendix 2). The fields most needing explanation are "champion" and "backwards compatible".

ALTO board members are volunteers, and, as already mentioned, have full-time work elsewhere. Asking a board member to carefully study one or two proposals of the dozen or so proposals before the board is much more doable than asking him/her to study all of them. Hence, in order to consider proposals efficiently and to reduce the demands on board members' time, each proposal is adopted by a board member who volunteers to "champion" it. The champion studies the proposal, considers its implications for the current ALTO schema, especially on backward compatibility, and, during the course of one or more meetings, explains to other board members why the proposal ought to be adopted or rejected. In other words, a proposal's champion is its advocate, but not the usual sense of the word *advocate*. The champion may also advocate for its rejection if the proposal is inappropriate.

Backward compatibility is most important for digital library software which uses ALTO. In order to use ALTO XML files which incorporate new features from a changed ALTO schema, the software may very likely have to change. If the new feature "breaks" some feature in the old schema, then it is necessary to have 2 different code paths to accommodate both new and old ALTO files[7]. "Breaking change" is something that software developers and digital preservationists strongly prefer to avoid.

Obviously backward compatibility is an important consideration with any schema change. It is preferable to deprecate old features in favor of an improved new feature since when a feature is deprecated, files produced with the old and new versions of the schema are compatible with each other (sort of). Deprecated features may be supported for one or two future current versions and thereafter completely phased out.

[6] IMPACT, or Improving Access to Text, was an EU-funded project whose objective was to "significantly improve access to historical text and to take away the barriers that stand in the way of the mass digitisation of the European cultural heritage" (cf. http://www.impact-project.eu/). The IMPACT project has been succeeded by the IMPACT Centre of Competence whose goal is to "make the digitisation of historical text better, faster, cheaper" (cf. http://www.digitisation.eu/).

[7] It may or may not be possible to transform ALTO files conformant to the old ALTO schema into files conformant to the new schema.

If backward compatibility is preserved in a new ALTO schema, the schema will be released as a minor versions, for example, version 2.1. If, for some reason, backward compatibility is broken by a new schema, a major version will be released, for example 3.0.

New schemas will be released as needed but no more than twice in a year and on a schedule which matches new METS schemas (January and July). A draft schema will be available for public comment at http://www.loc.gov/standards/alto/ one month prior to its release as a new major or minor version.

The ALTO Editorial Board has a wiki for meeting agendas, meeting minutes, working documents, and change proposals (http://altostandard.pbworks.com/). The public is welcome to join the wiki and to comment on current change proposals. One can request access on the wiki homepage.

The ALTO board believes that a stylistically uniform schema will facilitate the understanding and use of ALTO so it has drafted and will soon formally adopt design principles (see Appendix 3).

5. Meetings

As mentioned above, meetings are mostly by teleconference or web conference. Occasionally, as work schedules and, even more importantly, as employer budgets allow, the ALTO board meets face-to-face, and always in conjunction with another library conference such as the DLF Forum or the IFLA World Library and Information Congress.

By consensus the board meets at 2pm UTC on a Thursday. There has been some attempt to settle on a particular Thursday of the month, for example, the 1st or 2nd Thursday, but member schedules have proved too variable for this. The date of the next meeting is decided during the current meeting (preferably) or by email (if needs be) or by Doodle poll (last resort).

One of the board members (currently Frederick) assembles a draft agenda prior to a scheduled meeting and emails it to board members via the ALTO listserv. The email with the draft agenda asks other board members for additional agenda items for the next scheduled meeting. The draft agenda is also posted to the ALTO wiki. The final agenda as well as the URL to the minutes from the last meeting are emailed to ALTO listserv members a couple days prior to the scheduled meeting.

The board has tried plain teleconference, Skype, Skype with desktop sharing, and Webex for its meetings. In principle either Skype or Skype with desktop sharing is attractive because Skype is so widely used and free or inexpensive (Skype desktop sharing isn't free). But both have proven to be extraordinarily unreliable. Fortunately the employer (CCS) of one of the board members has a Webex subscription. Webex is very reliable, and because it allows desktop sharing, it also gives the meetings both a visual and an audio channel. In future we may try Google Hangouts in order to remove our dependency on a fee-based subscription service.

We recommend that members call from a quiet place or, if this is not possible, to mute their microphones when they are not speaking. Background noise, like keyboards typing, music playing, or officemates talking, can make it very difficult to hear and understand what's being said.

One of the board members (currently Frederick) moderates the meetings. All members are encouraged and expected to contribute, sometimes extemporaneously and at other times after prior preparation. For example, one of the current outstanding action items (see below) which does require preparation outside of a board meeting is to create design principles to guide future changes to ALTO.

Most meetings produce one or more action items. Each action item is assigned to one or more board members. An action item does have a start date, the date on which the item was created, and may have a completion date, for example, "by the next meeting". But since board members are volunteers with full-time "other" employment, no one is chastised for missing a deadline. Spoken or unspoken commendation by one's fellow board members and the knowledge that one is being of service to the library community is the only positive motivation while the only negative motivation is the wish to avoid the embarrassment of disappointing fellow members.

6. Board Member Motivations

What motivates board members to join and participate? All members are full time employees of other organizations and presumably have plenty to do for their employer. Board members are unpaid volunteers and must therefore be intrinsically motivated. According to Wikipedia

> *…intrinsic motivation refers to motivation that is driven by an interest or enjoyment in the task itself, and exists within the individual rather than relying on external pressures or a desire for reward. [People] who are intrinsically motivated are more likely to engage in the task willingly as well as work to improve their skills, which will increase their capabilities.*[8]

This is obvious to anyone who has volunteered to do a non-trivial task for a demanding, but uncompensated, position or to someone who has managed a volunteer organization. It is nevertheless important to keep in mind both board members' fulltime work and intrinsic motivations: Both factors contribute to member's independent observations, perceptions, and opinions.

Perhaps the question of motivation is best answered by some members themselves:

> *The Library of Congress has a strong interest in maintaining library standards in general, and digital standards are particularly important, give the ever-changing nature of the medium. I serve on the Board to ensure that changes keep pace with technology but also retain functionality for the large body of existing data from years of scanning efforts.*
>
> *Nate Trail, Library of Congress, Washington DC USA*

> *Bibliotheque nationale de France has used ALTO from the very begining of its digitalization projects, and it now has millions of ALTO pages available for preservation and diffusion purposes. ALTO is a great tool used everyday,*

[8] Wikipedia contributors, "Motivation," Wikipedia, The Free Encyclopedia, http://en.wikipedia.org/wiki/Motivation (accessed Jul 2013).

everywhere. But ALTO also has a future, and the ALTO board is the right place to build it.

Jean-Philippe Moreux, Bibliothèque nationale de France, Paris France

The Singapore National Library Board (NLB) uses extensively the ALTO standard for its popular NewspaperSG service. The ALTO editorial board provides me the opportunity to meet and work with members with substantial experience with the ALTO standard and implementations.

Kia Siang Hock, Singapore National Library Board, Singapore

The Koninklijke Bibliotheek (KB) began digitizing printed material on a large scale around 2005. Shortly after that ALTO was chosen and is still used as an important part of the format the KB has designed for the now many millions of pages digitized material and growing. In the future we hope that it will also be possible to improve the quality of the digitized collection, for example, the quality of the text. For these reasons the KB as well as I are interested in helping the community to maintain and develop the standard.

Evelien Ket, Koninklijke Bibliotheek, den Haag, the Netherlands

Since 2000 I've been creating digitization workflow software and or managing text digitization projects of all sizes. As an ALTO board member I have the opportunity to influence the future direction of one of the principle standards used in text digitization. Besides, if one belongs to a community, one has an obligation to contribute to it.

Frederick Zarndt, IFLA Newspapers Section, Coronado CA USA

7. Conclusion

As a standard ALTO has had an interesting life, coming out of a joint academic / commercial project, being maintained for a while by a particular company (CCS), and then returning to a public, open standard. It has a proven track record and millions of documents are scanned and expressed using it, ensuring that it will continue for many years. Paired with METS, this standard ensures that digitized paper documents can be electronically understood with precision and clarity. The Board is committed to maintain it's usefulness into the future.

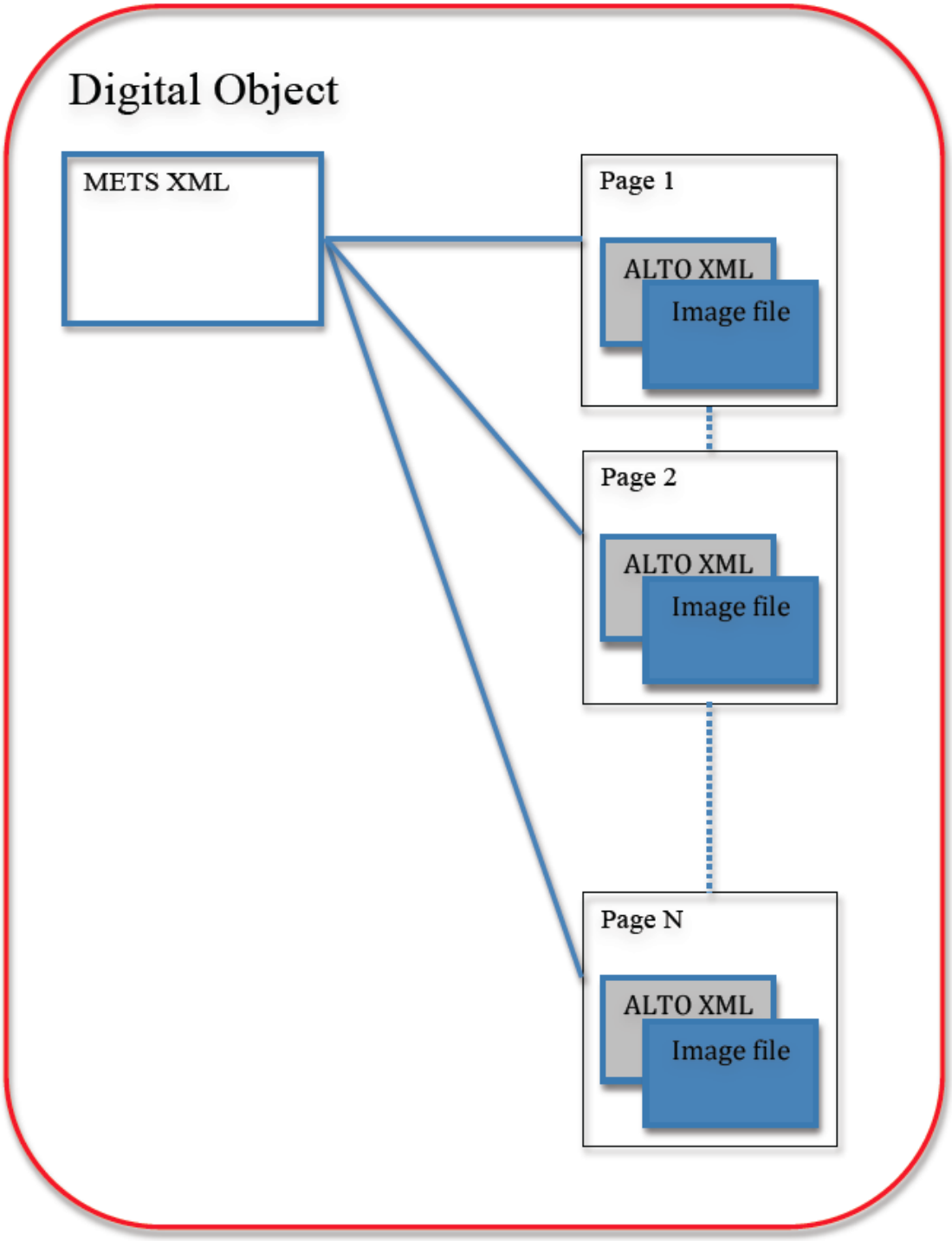

Diagram Showing Use of METS and ALTO XML Files to Represent a Text-based Digital Object such as a Book, Magazine Issue, or Newspaper Issue

Appendix 1: ALTO Editorial Board Membership Criteria

1. The ALTO Editorial Board maintains editorial control of ALTO, its XML Schema, and official ALTO documentation. Additionally, the Board promotes the use of the standard and endorses best practices in the use of ALTO as they emerge. The ALTO Editorial Board is representative of important communities of interest for ALTO.
2. Board member criteria:
    a. Have significant experience with ALTO implementation or ALTO-related tool building either previously or currently
    b. Represent either currently or previously one or more of the following constituencies from the international digital library community:
        i. National, Academic or Public Libraries
        ii. Information services utilities
        iii. Governmental agencies or organizations
        iv. Vendors supporting digital library operations
    c. Demonstrate experience in one or more of the following areas:
        i. XML or other information encoding languages
        ii. Digital library or digital repository implementations using ALTO
        iii. Metadata for digital libraries or repositories such as descriptive, administrative, structural, or transport schemas
        iv. Tool development and /or use for digital information creation, capture, storage and management, discovery or retrieval
    d. Committed support from home institution to support telephone and web conferences calls, face to face meetings when feasible, in-person or online training events, and other Board activities
    e. Demonstrate ability and interest in developing and fostering the use of ALTO within digital libraries / repositories and building a strong ALTO community of implementors
    f. Ability to commit to a 2 year term with the possibility of renewal.
3. Expectations for ALTO Board member participation:
    a. Makes a serious commitment to meeting the Mission & Objectives of the ALTO Editorial Board.
    b. Participates actively in the work of the Board to maintain and promote the ALTO schema.
    c. Regularly and actively participates in periodic telephone and web conference calls, as well as face to face meetings when feasible.
    d. Prepares for meetings, stays informed about committee and work group activities, and reviews and comments upon meeting notes, committee and work group reports, and ALTO communication media as appropriate including the ALTO listserv and ALTO wiki.
    e. Gets to know other Board members and builds collegial relationships among Board members that contribute to informed and congenial decisionmaking.
    f. Exercises professional judgment about changes to ALTO and the impacts of changes upon current implementations as known by personal experience or by input from other ALTO implementors.
    g. Participates in committee or work group activities, educational training events and other ALTO promotional and fundraising activities.
    h. Commits to a 2 year term of appointment with the possibility of renewal.

Appendix 2: ALTO XML Change Proposal Template

Champion	board member name
Submitter	submitter name and email
Submitted	YYYY-MM
Status	**submitted / discussion / review / accepted | rejected / draft / published** submitted - initial status when proposal is submitted discussion - proposal is being discussed within the board review - xsd code is being reviewed accepted - proposal is accepted rejected - proposal is rejected draft - accepted proposal is in public commenting period published - proposal is published in a schema version
Backwards compatible?	**UNCLEAR / YES / NO**
ALTO version	version where proposal will be included

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam ac ultricies augue. Pellentesque consequat interdum nulla, placerat ultricies magna scelerisque nec. Phasellus eget pellentesque magna. Phasellus dolor leo, vulputate ut tempor sit amet, ultrices eget turpis. Ut ornare convallis euismod. Aenean convallis elit feugiat augue dapibus pharetra. In at leo purus. Fusce faucibus iaculis orci, a luctus augue ullamcorper quis. Nulla magna nibh, elementum ut pellentesque non, fringilla sagittis nulla.

Example

```
<Glyph ID="P4_ST00001_G02" CONTENT="2">
  <Shape>
    <Rectangle HPOS="240" VPOS="223" WIDTH="10" HEIGHT="24"/>
  </Shape>
  <Variance CONFIDENCE="0.5">s</Variance>
  <Variance CONFIDENCE="0.1">8</Variance>
</Glyph>
```

Current Schema ALTO 2.0:	Proposed change:
<xsd:complexType name="processingStepType"> <xsd:annotation> </xsd:annotation> <xsd:sequence>	<xsd:complexType name="processingStepType"> <xsd:annotation> **<xsd:documentation>A processing step.</xsd:documentation>** </xsd:annotation> <xsd:sequence>

Appendix 3: ALTO Schema Design Principles

INTRODUCTION

The purpose of these principles is to provide guidance for future development of the ALTO schema. The document defines naming-rules for elements and attributes so that a coherent name-style will be used over time.

General

- Purpose of an element must be unambiguous; Information that are expected to be stored in an element must be defined clearly as part of the schema documentation.
- Ensure integrity of the data; Information that have the same(!) semantics should only be stored once in an ALTO file to reduce file sizes and ensure integrity.
    - References between XML elements should be established case using ID/IDRef mechanisms
    - XML Elements should be nested to represent a “comprises of” relationship between real-world objects that are represented by the elements
    - Information that qualifies the value of an element should be recorded as an attribute of the element itself. This may e.g. include encoding information etc.
- The ALTO-Schema should be used as a stand-alone schema and not borrow elements from other namespace. Therefore every ALTO-document must define the ALTO-namespace at the root-element level; other name space declarations on embedded elements are not allowed.

Elements and attributes

- Names of XML elements and attributes must only contain ASCII letters.
- Names of XML elements and attributes shouldn’t be longer than 20 characters.
- All names must be Camel-Case.

Schema design

- Elements (<xsd:element>) should be defined as global element in the schema. Global elements are elements that are direct descendants of the root element of the schema.
- Global elements should be re-used in the schema instead of defining local elements.
- Global elements must not have different names when they are re-used, except when they are extended and a new element is derived.
- Cardinality of elements should be expressed explicitly in the schema (using minOccurs and maxOccurs).
- The ALTO schema should not be modularized

Specific Principles

The current ALTO schema holds two different information objects:

- administrative metadata about the file and its provenance
- full text: the actual full text with layout information as well as provenance information of the full text itself

The use scenarios for both metadata types are different. Administrative metadata is very rarely being processed. It is just stored with the ALTO file in the repository and usually not queriable. Unlike the administrative metadata the full text is queriable. It is stored in and accessed from various systems: retrieval, exchange and render-systems.
Therefore the design requirements changes

Specific requirements for Administrative Metadata

- Mixed content elements (text and child-elements) must be avoided
- The order of elements should be enforced wherever possible (using xml-schema's sequence compositor)

Specific requirements for Full Text

- All changes that are made should be backward compatible so that ALTO files that comply to an old version of the schema will also comply to the new version of the schema.