TED (TEmplated Database)
User Guide

Purpose

This guide describes the TED (**TE**mplated **D**atabase) Service for prospective users. Information is included on the scope and coverage of TED, as well as the general steps and decisions which must be made in order to convert or create a database for a collection in TED.

Read this Guide in its entirety to determine whether or not your collection would be appropriate for TED. If you wish to proceed further, follow the steps in the Getting started on your TED project section.

Contents

May 8 2008

http://hul.harvard.edu/ois/systems/ctools/ted-userguide.pdf

1. What is TED

Librarians and archivists have long lamented the fact that much current research is in danger of being lost because it is sitting in discreet collections on a computer under a desk in someone's office. Not only is the data inaccessible, there is little hope that the data can be preserved. What is needed is a system that can take an arbitrary data layout, store it in a format that enables preservation and provide access to the contents via the web. The TED (**TE**mplated **D**atabase) service has been developed to meet this need.

Using XML technology, TED allows a curator to create a standard database for a collection, define a structured data format, and easily customize screens and parameters for search and display. This data can be maintained directly in TED or batch loaded from an external source. TED offers a web-based public interface that enables end-users to search, browse, view result sets, and retrieve records for items in a collection, as well as link out to digital objects when such links are provided.

Components of TED include an import/loading facility with data validation, data storage, a web-based maintenance interface, and a public web search and browse interface. The collection search interfaces can be customized with regard to text, images and search delimiters to provide each collection with a unique public identity.

The intent of TED is to facilitate web access to collections without requiring an extensive OIS implementation effort or the need for high-level programming skills. In general, collection information that can be represented as XML documents has the potential to be added to TED. In other words, TED is designed to work as a metadata catalog containing records (XML instance documents) that describe individual items (or groups of items) in a collection. In some cases, TED might also be used as a collection database – a repository for the digitized content of a collection (as long as the content can be represented by text in XML). See the TED record format section for more information.

TED is proposed to be used for collections that do not fit within the scope of other existing Harvard catalogs, and yet do not have sufficient size, audience, or topical generality as to warrant the creation of a new customized cataloging system. Instead, TED provides a generic system capable of simultaneously hosting a heterogeneous set of collections.

TED was developed by the HUL Office for Information Systems as part of the HUL Library Digital Initiative (LDI).

2. Eligibility to use TED

2.1 Who can use TED?

Any Harvard organizational entity is eligible to use TED. Harvard organizations outside the library system are eligible if sponsored by a Harvard library. Individual members of the Harvard community are also welcome with the sponsorship of a Harvard library.

Harvard library sponsorship is required to ensure a long-term commitment to support a collection once it is available in TED. In the event that the original collection owner (not a library) can no

longer support a TED collection, a library sponsor agrees to ongoing financial and curatorial responsibility for the collection.

2.2 Is your collection appropriate for TED?

The primary requirement when considering TED is that a collection consists of library-like materials with long-term intellectual or curatorial value. Additional scope and content guidelines are listed below. Please also review the Components of TED section to get a sense of whether TED functions will meet your needs.

Scope of TED

Note that collections cataloged in TED can belong to any academic discipline, subject domain, etc. There are no limits on the types of elements or vocabulary used. TED may be an option for:

- Collections that are not appropriate for other existing HUL Union Catalogs, such as HOLLIS (bibliographic data), VIA (cultural heritage images), OASIS (manuscript and archival finding aids) and HGL (geospatial data).
- Collections which are relatively small and targeted to specific audiences and topics.
- Collections with relatively static use for which public web access and data preservation are the primary goals. TED is not designed for transactional systems which require a significant amount of ongoing data processing.
- Collections that do not require a full-featured collection management system.

TED content guidelines

- **Use TED as a catalog or a collection database.** The information stored in TED can be metadata (cataloging) that describes digital and non-digital objects in your collection (a TED catalog), or it can be the actual data that makes up your collection (a TED database). Presently, there is no size limit on this data. See Harvard collections using TED (on the OIS web site) for examples of how TED is being used.
- **TED can contain text and thumbnail images.** Data types that can be stored in TED include: text files (encoded in XML) and thumbnail images (JPEG or GIF). If data in TED includes links to digital objects other than thumbnails, these objects must be stored external to TED (for example, in the Digital Repository Service or a local network-accessible file system).
- **Character set support.** TED supports the same set of characters as the HOLLIS Catalog. Currently this includes the Latin-based scripts used for Eastern, Western, and Northern European languages, including diacritic marks, punctuation, ligatures, special characters, etc. Read more about TED character set support.
- **TED data must be publicly accessible.** Data stored in TED must be publicly accessible and cannot be restricted to Harvard users or a subset thereof.

- **Collection data can be created within TED or loaded from an outside source.** See the Import/loading facility section for the details.

3. Components of TED

Components of TED include an import/loading facility with data validation, data storage, a maintenance interface, and a public web-based search and display interface.

3.1 Import/loading facility with data validation

Data can be batch loaded into TED via FTP, or created online via the TED web maintenance interface. More information about the TED web maintenance interface appears later in this document.

Batch loading is used to populate TED with the initial set of collection data but it can also be used for ongoing maintenance (in cases where the collection owner chooses to maintain data in a local system and periodically transfer this data to TED).

The batch load process supports addition, replacement, and deletion of individual TED records (called XML instance documents). During a batch load, incoming documents are matched against the unique identifiers of existing documents. (As part of the collection set-up, you will specify the unique document identifier field for your collection.)

XML instance documents that are created from scratch in XML editors other than TED should be validated against the XML Schema for your collection and the general TED XML schema.

See Section 4 for more information about the TED XML schema. The batch load into TED will also validate the documents against the Schema. The validation will include a check for such items as tagging accuracy (all the begin-tags and end-tags must match up), mandatory fields, what tags can appear in the document and how they must nest within each other, etc.

3.2 Data storage

The information stored in TED can be metadata about digital and non-digital objects in your collection (a TED catalog), or it can be the actual data which makes up your collection (a TED database).

TED data types

Data types that can be stored in TED include: text files (encoded in XML) and thumbnail images (GIF or JPEG). If data in TED includes links to digital objects (other than thumbnails), these objects must be stored external to TED. For example, in the Digital Repository Service or a local network-accessible file system. Presently, there is no size limit on this data.

TED character set support

TED supports the same set of characters as the HOLLIS Catalog. This is the MARC-21 repertoire, which includes characters defined by ASCII, ANSEL (Extended Latin), Greek, superscripts and subscripts, Hebrew, Cyrillic (basic and extended), and Arabic (basic and extended), and the NISO Z39.64 East Asian characters.

This support is being implemented in a phased manner. Currently, TED supports the Latin-based scripts used for Eastern, Western, and Northern European languages, including diacritic marks, punctuation, ligatures, special characters, etc. (Note that this excludes Greek and Cyrillic.)

Current language coverage includes: Afrikaans, Breton, Basque, Catalan, Croatian, Czech, Danish, Dutch, English, Esperanto, Estonian, Faroese, Finnish, Flemish, French, Frisian, German, Greenlandic, Hungarian, Icelandic, Irish, Italian, Latin, Latvian, Lithuanian, Maltese, Norwegian, Polish, Portuguese, Provencal, Rhaeto-Romanic, Romanian, Romany, Sami, Slovak, Slovenian, Serbian, Spanish, Swedish, Turkish, Vietnamese, and Welsh.

TED will follow Aleph in its implementation of non-Roman scripts (Greek, Cyrillic, Hebrew, Arabic, East Asian languages, etc.).

3.3 Maintenance System

There are two options for maintaining the records for your TED collection: batch loading records from a separate source database or creating/modifying records in TED using a web-based maintenance interface.

Batch load records from a separate source database

You can maintain your data in a separate source database and periodically batch load new and changed records to TED. Remember that by policy, all TED collection records must be publicly accessible (and individual TED records cannot be suppressed). If the local source database contains records that should not be made public, these must be excluded from the batch deposit process.

Create/modify records using the TED web maintenance interface

The TED maintenance system enables an authorized user to create, modify, and delete XML records (instance documents) via the web. However, this interface is not designed to be a full-featured sophisticated system for database and collection management. The TED maintenance system currently has the following features:

- A **customized user interface** according to the fields and layout of your collection records. For new records, a web form will be displayed which will allow you to input data for each field in the collection schema. If you are editing a record, the current data in each field will be displayed for you when you call up the record.

- **Flexible control over reusable data elements.** Define one field or a group of fields so that a data value entered in one field is available in a drop-down list for other fields. This feature (defined in your collection's XML schema) can reduce keystrokes by eliminating some original

keying of data. For example, if there are many contacts with the same location and address, you will only need to enter that location and address one time.

- **Required fields and validation.** The system will validate fields indicated in your schema as containing only certain XML data formats, such as integers, dates, decimals, alpha characters, etc. Also, you can define within your collection schema that certain fields are required.

- **Partial records are not supported.** This means that all required fields in the record must have values before you exit the system.

- **No record suppression.** Individual records cannot be suppressed from public view, but individual fields within records can be selected as non-displaying when setting up the XML schema for your collection.

- **No support for authority control** in record creation and maintenance operations.

- **Limited support for global search and replace** operations within a set of documents.

3.4 Public web search and browse interface

TED will enable users to search and browse data, view result sets, display records for items in your collection, as well view search history and download records. Collection records can link out to digital objects that are stored external to TED (this is optional).

Consult the TED Customization Guide for information on public interface functions and customization options.

Search and Browse

Search. TED supports simple and advanced search. In **simple search**, the user can search for individual terms (implicitly ANDed together) or a quoted phrase that occur anywhere across all records in a collection. **Advanced search** also offers Boolean AND, OR, NOT, and searches can be limited to a specific subset of fields. You can specify the fields and parameters for simple and advanced search in your database and also select fields and data elements which can be used to limit other searches.

A wildcard operator (*) can be specified at the start, end, or intermediate position of a word, e.g., *plated, templat*, *plat*. The user will be able to specify on the search screen the number of results to be returned per page.

Browse. The TED Browse option allows users to browse the contents of selected indexes which you specify for your collection. Users can also jump to alphabetical points in the indexes.

Both search and browse will return the number of hits and a list of results. Each entry in a results list contains brief information about the record and a link to display the full record. From the results list it is possible to save individual records for downloading.

Note that searching across separate TED collections is not a supported function. There are plans to implement a cross-catalog search function, using a product called Metalib (by Ex Libris).

Display

The full record display includes all fields in a record that you have defined as publicly accessible. If your collection's XML schema defines a record as having hierarchical sections and subsections, the full record display will provide a means to visually expand or contract each level of the record. If your data includes thumbnail images, these can appear at any point in the record. There may also be links within a record that point to external digital objects (stored in the DRS or another network-accessible file system).

From the full record display it is possible to save the record to a set for downloading. It is also possible to mark selected sections of the full record so that only these are included when the saved record is downloaded.

Save/History

The TED save function (also called the TED “clipboard”) provides a means of grouping selected records for export via email or other delivery methods. Records may be added to and removed from the clipboard freely.

TED also provides a search history restricted to the current browser session.

4. TED record format (XML schema)

TED uses XML technology as the basis for describing, storing, searching and retrieving data about collections (metadata), or collection data itself. XML (E**X**tensible **M**arkup **L**anguage) is a widely used markup language for documents containing structured information. It is called extensible because unlike HTML, XML supports flexible and adaptable identification of structured information.

Most documents have structure. Structured information contains both content (words, pictures, etc.) and some indication of what role that content plays. For example, content in a section heading has a different meaning from content in a footnote, which means something different than content in a figure caption or content in a database table, etc. Almost all documents have some structure.

An XML schema describes structure of documents in a collection. The XML specification defines a standard way to add markup to documents. In other words, XML provides a facility to define tags and the structural relationships between them. A **schema** is a formal description of a document structure, in particular of a structure expressed in XML. XML schemas express shared vocabularies and allow machines to carry out rules made by people. They provide a means for defining the structure, content and semantics of XML documents.

XML "instance documents" are individual records that follow a collection schema. In order to create a customized catalog or database for your collection, the data vocabulary for your collection will be defined in an XML schema, and documents that comply with this vocabulary, called **XML instance documents**, will be created and stored in an XML database. As a best practice, it is recommended that collection schemas include up to three levels of hierarchy, with a maximum of thirty distinct elements per level, each of which may include optional or repeatable sub-elements. However, this will vary with the collection. One of the data elements that will be required is a unique record identifier for each document in your collection.

An XML schema also controls public display of data. It is important to understand that the design of your TED schema will affect not only the metadata content but also the way that the records are displayed in TED. The order and nesting of elements in the schema will correspond to the order and nesting in the display.

Your TED liaison will work closely with you and the OIS metadata analyst to identify the appropriate data elements and develop an XML schema for your data which will be appropriate for storing, searching and retrieving documents and linked items in your TED catalog. They will also work with you to map data in order to convert existing formats for ongoing collections.

Links to XML schemas are available in the "Documentation" section of the TED website: http://hul.harvard.edu/ois/support/docs-ctools.html.

5. Security of data in TED

The web-based public interface to a TED collection will be freely available with no access restrictions. Access to digital objects linked-to from a TED collection record may be restricted based on the policy of the repository in which the object is stored. All aspects of access to external digital objects are outside the scope of TED.

The TED maintenance system is only made available to collection managers and their designated staff. Access requires a Harvard ID and PIN registered to use TED maintenance. Authorized users will be assigned a specific role that controls the types of maintenance tasks they can perform. Available roles include:

- superuser (create, modify, delete in a particular collection)
- admin (create and modify in a particular collection)
- staff (modify in a particular collection)
- public (read only in a particular collection)

The collection manager will be set up as a superuser in TED and all of the other roles are optional. Appropriate OIS staff may also have superuser access privileges to the TED database as necessary.

The TED section of the OIS website has instructions for requesting TED maintenance access.

Access to TED maintenance: http://hul.harvard.edu/ois/systems/ctools/maint.html.

6. Billing for TED

Fees will be charged for each collection in TED to allow OIS to recover a portion of the costs associated with the original set-up and configuration, advisory services, data conversion, and ongoing maintenance; these fees will vary with the complexity of the project. Fees will include a one-time start-up cost and an annual license fee of $520. There will be one of three levels of start-up cost assigned to most TED projects based on complexity:

Basic - $2,000

Medium - $4,000

Complex. - $8,000

There will normally be no charge for data storage of TED catalogs or databases.

Other charges may need to be assessed, including costs for additional programming, special or complex data conversion, excessive changes to specifications, or other exceptional issues.

Your TED Liaison will work with you to determine billable costs, and you will be asked for a financial contact and billing code.

7. Getting started on your TED project

7.1 Requesting participation

The first step toward participation is to determine whether TED is a suitable environment for your collection data. To do this, the collection owner reviews TED requirements, submits a questionnaire describing the collection, and then meets with OIS staff to evaluate the collection's eligibility for TED. Full instructions on requesting participation are available in the TED section of the OIS web site:

How to participate: http://hul.harvard.edu/ois/systems/ctools/overview-ted.html

An OIS staff member will be assigned as your **TED liaison** to help you prepare for the TED project approval process. Proposals that receive a positive initial review are then passed to OIS for coordination of the final evaluation and approval process..

Your TED liaison will also work with you to determine a timeframe for setup of your collection in TED. The timing of each TED implementation project will vary based on the complexities of the collection (the data, conversion processes, user interface configuration, etc.), use of other HUL infrastructure systems (DRS, NRS, etc.) and availability of OIS technical staff. It is likely that any TED implementation project will take a *minimum* of three months (from start of setup activities to "go live").

7.2 Project setup checklist

Once a TED project gets approval, a variety of setup activities begin. The collection's Project Unit Manager will work with the TED Liaison and other OIS staff to accomplish a series of project milestones (noted below).

A. Selecting a unique record identifier

You must select a unique record identifier for each document in your collection. This can be an existing identifier (for example, a local accession number or database system number). Record identifiers can consist of any number of alphanumeric characters.

B. Developing a record format (XML document schema)

Your TED Liaison will work with you to identify the set of data elements that will be present in individual collection records. If your data is hierarchical in nature, then data elements at each level need to be identified. Read more about the TED document schema/format.

You will then work with the OIS Metadata Analyst to develop an XML document schema that incorporates these data elements.

To view the TED collection XML schema, look here:
http://hul.harvard.edu/ois/xml/ns/ted/TEDSchema.html

To view a sample collection XML schema, look here:
http://hul.harvard.edu/ois/systems/ctools/sampleXMLschema-parry.html

Checkpoint: you will be asked to sign-off on the final XML Schema once it has been approved by the OIS Metadata Analyst. Changes to this schema after it has been finalized may incur additional setup charges and delay implementation of your collection in TED.

C. Planning data conversion

If you plan to migrate existing data from another system to TED, a conversion specification will be needed. You will work with your TED Liaison and the OIS Metadata Analyst to develop a conversion specification that maps fields from your existing collection records into the TED XML document schema.

Actual data conversion can be performed by OIS or by the collection owner. Conversion performed by OIS is included in the initial setup fee *unless* circumstances require development of special or complex conversion processes. More information about fees can be found in the Billing for TED section.

D. Customizing the public interface

The TED public interface must be customized to get a "look and feel" that is suitable for your collection. You will need to make decisions about general display options (top banner, navigation links, page footer, textual instructions), parameters for search and browse options, and content/layout of record displays.

You will record your customization choices using instructions and page templates found in a separate document called the *TED Customization Guide* (available from the TED documentation page). Normally, customization of the TED public interface should begin **after** you have made substantial progress in defining your collection's XML schema. Your TED Liaison can answer any questions that arise as you complete the *Customization Guide*.

E. Linking to your TED Collection

Links to your TED database can be included in any or all of the following: a collection-level record in the HOLLIS Catalog (strongly encouraged), the Harvard Libraries Portal website, one or more appropriate organizational Harvard websites, a website specific to the collection, a cross-catalog search of all TED catalogs, etc. It is a requirement that all TED databases be publicly available and not restricted to the Harvard community or a subset thereof.

F. Testing data loads into TED

Once the XML document schema for your collection is developed and then approved by the OIS Metadata Analyst, and the public interface is customized for your collection, OIS will load several records into a test version of your TED database so that you can see the records displayed and test the search, browse and navigation features. The batch load process will also be tested. Adjustments will be made as necessary until the results are satisfactory.

G. Loading your data (officially)

If you will be migrating data from a local system into TED, the last step will be the final production load of converted data. OIS will run a final batch load of data into TED to create your initial database.

Checkpoint: you will be asked to review this data carefully and signoff on the data load. Problems missed during review that require re-conversion/re-loading may incur additional setup fees and delay implementation of your collection in TED.

H. Creating and maintaining new data in TED

Once existing data is loaded into TED, you will be able to start creating and modifying data directly in TED. If you plan to maintain your data in an external system, you will be able to start batch load updates to TED.

I. Finalizing payment schedule and fees

OIS will work with you to determine the final price for your TED project and a mutually agreed-upon billing and payment schedule.