Slide 1

A couple of points: this presentation and my notes for it are in the UO's institutional repository under the URL (or handle) listed on this screen – it's also on the first page of your handout. I have also supplied a one-page listing of some references that I either refer to or consulted in preparing this.

My first disclaimer: I am not an expert on metadata, nor am I a theoretician. I am primarily a self-taught practitioner who has tried to keep up with the changes in the profession by reading and attending lectures, workshops, and conferences. I'm sure that some of you in this audience are much better versed in metadata theory – and practice – than I am.

The other disclaimer is that I will be discussing metadata within the context of two content management systems: DSpace and CONTENTdm. Although both have led us to make some decisions that we would rather have handled differently, we have been very satisfied with them and they have enabled us to accomplish many things that we wouldn't have been able to do otherwise. For us, their advantages have far outweighed their current limitations.

Slide 2 Acknowledgements

I'd like to acknowledge the creativity of a member of my department, Marion Obar, who came up with the slogan I've used as the title of this presentation and who gave me permission to steal it.

Slide 3 UO's Digital Collections Home

While the challenges of creating and maintaining digital collections are many, I'm going to focus on the collection and maintenance of metadata and how easy it is to go wrong.

The University of Oregon Libraries, like many libraries and other cultural heritage institutions, began to create and provide access to a variety of digital content about a decade ago. Some of the earliest efforts began in Special Collections & University Archives with the digitization of portions of the library's holdings of more than 400,000 rare photographs. Scanning and metadata collection were done with little awareness of, much less adherence to, emerging standards. Digital files were stored on a variety of media, with no systematic backups, no checks for file integrity or media degradation, and little internal consistency in file naming, creation of digital images, or collection of metadata. In the space of a few years, some files were no longer usable, not just due to degradation of storage media but also due to inadequate collection and tracking of metadata.

Slide 4 Metadata Implementation Group

In 2002 we established a Metadata Implementation Group to develop the standards to be applied for digital collections that we were going to build using CONTENTdm software.

In March 2003, we identified a collection of materials for digitization and mounting in CONTENTdm®. We created and made this collection publicly available in record time (four months) to fulfill our obligations to a granting agency. We achieved this by calling on the expertise of four librarians in what was then the Catalog Department, working closely with staff in Special Collections & University Archives and recruiting volunteers from the Catalog Department to learn how to scan glass plate negatives, apply descriptive and technical metadata, and load the digital objects and accompanying metadata into the CONTENTdm® system.

While this was underway, the library also established an institutional repository, called Scholars' Bank, using DSpace software. I was involved in both efforts as the chair or co-chair of the groups working to implement the systems.

Slide 5 Metadata and Digital Library Services

In December 2003, the Catalog Department was re-christened Metadata and Digital Library Services (MDLS) in recognition of its expanded role of implementing and maintaining digital collections –in addition to cataloging and preserving analog materials. Within the library, no new classified or professional positions have been added –all work on digital collections is an add-on to other ongoing work.

Slide 6 Factors affecting selection of metadata

That gives you some organizational background. Now on to metadata.

There are a lot of different factors that will affect the type of metadata you use for your digital collections. These are some of the ones that we have taken into account at the University of Oregon and I'll go into each of these very briefly.

Slide 7 Metadata schema

In the theoretical world, there are a large and growing number of metadata schema that you can pick from. My simplistic definition of metadata that I use when talking to students is that it's any information you collect about anything that you put into a structured framework, following agreed-upon or understood rules. I'll refer you to Agnew, Greenberg, and Schottlaender in the references for some good explanations of metadata and metadata schema.

These are some of the metadata schema that librarians are most familiar with. Again, I'm not going to define these but I will spend some time discussing the use of Dublin Core and what we have found to be its strengths and limitations.

Slide 8 Content standards

Content standards tell you how to structure the values in each metadata field. Content standards establish and document controlled vocabularies and rules for populating each data element.

There are many other content standards but these are the ones that we have used at the University of Oregon in our digital collections.

It is the lack of agreed-upon content standards, as much as the metadata schema, that affects the usability of collections beyond the local level.

Slide 9 Software considerations

These are some of the software considerations that have affected our choice of metadata at the UO.

The software being used for describing and making digital collections available is not as well-developed as library catalogs. It has also been designed to support collections being built by organizations other than libraries, as well as by libraries.

In theory, there are always best practices for the content of your metadata fields. For instance, a commonly understood best practice for subject terms used in descriptive metadata is to base the terms on a controlled vocabulary list. However, the choice of a particular controlled vocabulary may depend as much on the software being used and its limitations or features as on the target audience or the nature of the materials in the collection.

Without the supporting system functionality, making use of controlled vocabularies that were developed for use within one particular tradition, i.e. the MARC format and online library catalogs, is challenging.

Slide 10 target audience

Your primary and secondary target audiences will also have a significant impact on your choice and application of metadata. The type of information you decide to collect about your digital objects, how you label the fields, how you develop the content standards for the fields are affected by your target audience – and also by the other people or groups you're collaborating with.

One of the characteristics of digital collections that we have observed is that they are highly collaborative. People who haven't traditionally been involved in creating and providing access to traditional library resources – the end users and collection curators – are often involved in the decisions about the metadata elements and how the fields are populated. And they often have very strong opinions about the metadata fields and input standards – even if they don't use that terminology.

Slide 11 How is it being created or supplied?

The methods of collecting or supplying metadata have had a great impact on the type of metadata we collect and how reliably we collect it. Human supplied metadata is great for descriptions, although it tends to be subjective and very time-consuming.

In some types of collections, such as institutional repositories, the general public is supplying metadata about the digital objects. This clearly has an impact on the type and nature of the metadata.

Machine-generated metadata is great for certain types of information, such as information about file formats, file size, data integrity checks, unique identifiers, etc. How do you know which of these machine-generated data elements are significant, useful, necessary? There are as many opinions about this as there are digital repositories.

And, even if you can generate it automatically, it may still require substantial human effort to associate it with the appropriate digital object in a content management system or to get it into the form that agrees with the standards you have chosen to follow.

Slide 12 Functions it serves

There are almost as many different ways of characterizing metadata functions as there are metadata schema.

These are some of the functions that we have identified but different sources use different terms and different definitions for the categories. I won't go into what each of these mean (to us or anyone else) but I only want you to be aware of how complex the topic is and how easy it is to get embroiled in detailed discussions about terminology and definitions.

For us, the most complex category of metadata has been preservation metadata. If you're familiar at all with the PREMIS model that has attempted to provide guidance on just what information is needed in order to preserve a digital object, you'll appreciate the dilemma. This is also in the references I've provided.

Slide 13 Dublin Core Metadata Element Set

I'll be talking a good bit about Dublin Core for 2 reasons: 1, the software systems we use map data elements to Dublin Core and 2, it is considered by a large community to provide a good mapping between different metadata schema and to provide a good basis for interoperability. And we want our systems and collections to link up eventually with other similar collections, just as our library catalogs have.

One of the easiest criticisms of Dublin Core is the fact that it has two fields for essentially the same function: contributor and creator. The definition for these two elements is virtually identical. Why, in a supposedly Core set of data elements, are there two elements with such high overlap of functionality and such lack of precision in the definition? I'll get a bit into some other problems with DC a little later.

Slide 14 DSpace

DSpace software, used to build Scholars' Bank at the UO, maps fields to Dublin Core.

DSpace software follows a decentralized submission model, whereby individual authors or designated representatives may submit files and fill in the metadata (author, title, keyword, etc.) on the submission template. Because the developers of the software expected each community of users to have its own standards for the type of metadata to be supplied they built in no support for controlled vocabularies. Lists of controlled terms must be consulted outside of the DSpace framework.

Slide 15 Scholars' Bank

This is the home site of the University of Oregon's institutional repository, Scholars' Bank.

Slide 16 Adding a new field to an item

Once an item has been entered, someone with collection administration authorization can go in and modify the data in a field or add or remove fields.

The full DC simple and qualified element set is available for use but using some of the qualified elements results in them being hidden from public view or does not result in them being indexed. For this reason, and because of the high level authorization needed to make such changes, we have not modified the default metadata input form. We squeeze all the metadata into the defaults.

Slide 17 Public metadata for DSpace

This is a typical "item" record in DSpace showing the labels for the metadata. It's roughly the equivalent of a bibliographic record.

Slide 18 DC Metadata for DSpace

This is the same metadata showing the Dublin Core mappings for the fields. These are the default mappings. You can tell that some of this metadata was machine-generated.

Slide 19 default submission form

This is one of the screens of the submission form. Note that subject keywords is the default (and hardcoded) field label here for subject analysis of an item and it is mapped to simple DC subject. There is no link to any kind of controlled vocabulary. The limitation of the software and our own limited staff resources have meant that we are very relaxed in our input standards for subject access in our institutional repository. This was a metadata decision based on resources and software. Not on any theoretical best practice.

Slide 20	Logical or useful presentation

DSpace software was developed to handle articles submitted by individual authors very well, but it was not designed expressly to handle entire serial titles. The inconsistencies of the original publications in designating issues, coupled with the limitations of the software have presented interesting metadata challenges.

Our application of metadata in our IR has often been determined by the type of display we wanted to achieve. In the case of some journal issues, the only way to get them to collocate correctly within the collection is to reproduce the journal title at the beginning of every title field and to assign volume numbering and chronology to many of them – even when the originals lacked such designations. ()

Unlike an online catalog, where issue numbers are captured in holdings or check-in records, the way that we handled this publication meant that each issue has a separate record and it’s necessary to capture issue numbers somewhere in that record.

We chose to include it in the title field because it helps with orderly sorting.

Slide 21	Chronological displays of issues

This newsletter, *From the Center,* lacked any numbering and used only seasonal designations.

In order to get these issues to display in chronological order,
() we supplied a number for each issue in that calendar year it represented. (unless there was only one issue in that calendar year)

Slide 22	Actual digital object

This is an actual issue of the newsletter and you can see it lacks any of the helpful numbering.

We hereby violated one of the cardinal rules of cataloging – we supplied metadata that was not on the original item.

Slide 23 Dissociation

This is a journal title where we've gone even further – to meet the clearly articulated needs of the target community. We have digitized an entire journal run and have created a separate file – and an item/bib record – for every article.

To make the articles display in the order in which they appeared in the original publication, we added the journal title, issue number, and page numbers at the beginning of the title field – followed by the title of the article. We then added a title.alternate field to index just the article title by itself. The usage stats we are getting for these articles justify, in my mind, the effort we put into this.

Our metadata practices in our IR are based on:

- Dublin Core and its limitations
- Dspace limitations
- Decisions of the community liaisons and if there are any standards they want applied
- Staffing resources
- And whatever we have to do to collocate, present, and locate items reliably

It very definitely is NOT cataloging. As one of my metadata librarians said the other week when reviewing a document I had prepared to explain our practices (to ourselves, our colleagues, and our users): "It ain't pretty, but at least we're honest."

Slide 24 OAIster

We have verified that our metadata is compliant with the Open Archives Initiative Protocol for Metadata Harvesting and we register the IR in appropriate registries, such as OAISter and the Institutional Archives Registry.

This screen shows a Scholars' Bank item retrieved through OAIster, where we've registered the archive. Our materials are findable on the open web and through various registries.

The registries just require that the metadata be OAI-compliant but they have no content standards. Until such time as there are widely-accepted content standards for digital registries, we have decided to be flexible in the way we apply our metadata – as long as we document what we've done and periodically review our decisions.

Slide 25 Documentation of practices

This is a page where we have documented our practices and the rationale for them. We continue to add to it and I periodically review it and update these documents. Some of these documents have been modified three or four times to reflect our changing reality.

Slide 26 Picturing the Cayuse

Now on to some of our CONTENTdm collections and some of the issues we've faced with them. This is the home screen for our very first digital collection that was a collaborative project between the library and the Confederated Tribes of the Umatilla in eastern Oregon. This very first collection raised incredible metadata challenges, related to the software, Dublin Core, our target audience, who was collecting metadata, and staffing resources.

This is a collection of selected images from about 7000 photographs owned by the UO libraries taken by the white indian agent for the tribes in the early part of the 20th century. One of the goals of the project was to provide an opportunity for native people to describe images of their ancestors in their own words – and to do it before all of the people depicted in the images died off.

Since this was our very first collection, and since it was complex, we collected a LOT of metadata. New collections being set up today are done pretty differently, although we would still make some of the same decisions if we were building this collection today.

Slide 27 Full descriptive metadata

This screen shows some of the descriptive metadata for one image. We have fields where the label begins with TCI. TCI stands for the Tamastslikt Cultural Institute and the data in those fields were supplied to us by the tribal people using their own input standards and controlled vocabulary.

TCI notes is mapped to DC Description and we also have the General Notes field mapped to DC description that includes the description supplied by my staff. It turned out that TCI staff were supplying their metadata after ours in most cases and they had a good laugh at some of our descriptions.

We have the TCI terms and classes fields which both map to DC subject and our own subject field that also maps to DC subject. Different content standards but mapped to the same DC element.

Slide 28 Including technical metadata

This screen shows a fairly complete record including a great deal of technical metadata. We went a little nuts with this – erring on the side of caution and following the just-in-case model of collecting information. The argument goes that we don't know exactly what kind of technical metadata we're going to need to preserve these items, therefore we need to collect everything in very fine detail. However, our ability to actually capture and supply that detailed metadata consistently has not turned out to be possible with our staffing limitations – and with the fact that new collections are being thrown at us all the time. So, there's reality and there's illusion. Theory vs, practice.

Slide 29 Metadata challenges for group projects

These are some of the challenges you'll face in group projects.

Setting up your own local collections presents a lot of metadata challenges and requires a lot of decisions. However, if you are contemplating any projects that require contributing your metadata to a shared site where there will be expectations on the ability to search across multiple collections (like in a union catalog), I urge you to spend the time on hashing out the metadata ahead of time. AND HERE'S WHY…

Slide 30 UO's WWDL

This is the UO's home page for the Western Waters Digital Library

This grant-funded project involved 11 or 12 other institutions. The metadata from all of our local collections are harvested and collected by a central server. We had been building a joint collection for a year before there was any serious discussion of metadata field definitions and input standards. The results were interesting and some after-the-fact cleanup was needed.

Slide 31 Browse by format

In our local site, we early on set up canned searches that allowed us to browse all of our documents, all of our images, all of our aerial photos and maps.

Slide 32 GWLA WWDL home

This is the home page for the union catalog for the WWDL collection.

Slide 33 Metadata challenges

Underlying CONTENTdm is mapping to Dublin Core –simple and qualified.
Dublin Core itself provides a lot of latitude in application of its standards,
which is why there are application profiles being developed for a variety
of communities.

With 11 or more participating sites, local decisions on a basic core element can affect the effectiveness of searching across collections.

All project participants agreed to follow the Western States Dublin Core Best Practices, listed in your references. The document gives considerably more guidance on how to map different types of information to various Dublin Core elements. However, at least in the version we used, there was still considerable latitude in input standards for various fields.

In addition, some participants were harvesting from legacy collections created without reference to those standards, resulting in inconsistent values being supplied in the same field.

We set up a metadata task force to help us try to resolve some of these challenges.

Slide 34 Application of metadata standards

Date.original and Date.digital was the first issue that the metadata group worked on because we considered it low-hanging fruit – in that the standards were clear and didn't provide latitude.

Nevertheless, the legacy data from some of the harvested collections didn't conform. It still isn't clear if all of the legacy collections went back and cleaned up their metadata to conform to the more clearly articulated content standards.

Slide 35 DC mapping and aggregated searching

One of the greatest challenges of aggregated searching is with Subject.

This slide shows the underlying DC mapping for subject within the UO's WWDL collection.

Note that there are five different fields mapped to DC Subject. In the local UO collection, we have built search interfaces that allow us to create more meaningful searches.

Slide 36 Local and customized search interfaces

Locally, we can create separate search interfaces for the different types of subject. So, the fact that we use both LC and TGM terms is not a problem in our local site. They are in separate fields and we have built search interfaces to reflect that. On the central site, however, such customization and separation of distinct types of subject data is not possible because …

Slide 37 No mapping to encoding schema

One of the challenges for the aggregated searching is that the software does not allow for mapping to encoding Schema.

CONTENTdm does allow for mapping to qualified DC but does not currently extend that to encoding schema. If it did, that would provide some assistance for libraries using different source vocabularies.

At the moment, there is considerable redundancy and contradiction between TGM and LCSH. In an aggregated search, this can produce confusing search results.

Slide 38 Inconsistent search results

If one searches on subject within the aggregated site on the TGM term “Afro-Americans” there is only one result. If one happens to search on African-Americans, the LCSH term, there are 15 results (fortunately still including the single result found with the TGM search.)

The Western States standards recommend only that libraries use a controlled vocabulary, but doesn’t specify either the source vocabulary or the way it should be utilized. We are all in compliance with the Western States document but have made very local decisions. Reaching consensus on subject fields would be extremely difficult – yet this has an enormous impact on aggregated searching

Slide 39 Type recommendations

This was a particular problem area for us – the use of DC type. Prior to getting this cleared up and making firm recommendations that went beyond the Western States recommendations, searching on images retrieved a lot of documents because people had been interpreting the jpg of a page of text to be an image, rather than text.

These are the recommendations made by the Metadata Task Force for the Western Waters Digital Library. To get the records to display properly, the owning institutions needed to go back and correct their metadata.

Slide 40 Advanced search

How to make use of this in searches is explained on the advanced search screen.

Slide 41 Browse all images

Now when we search on images we get images

Slide 42 Browse all text

And when we search on text we get text.

Slide 43 The Future

Perhaps the future is one where the world at large is supplying metadata- without even knowing it. Here’s a screen shot from my personal del.icio.us site which allows you to capture bookmarks and assign metadata to them to help you find them. Del.icio.us calls the terms you apply tags.

Tags – allows to you classify your sites and then you can bundle various tags together. You can assign multiple tags to a site. Once you start to collect sites, your list of tags appears for you to choose from – allowing speedier and internally consistent classification

If your site has already been bookmarked by someone else registered with delicious, frequently used tags for that site also appear for you to choose from.

Slide 44 Folksonomies

This gets into what are being called folksonomies. Folksonomy is a word that combines "folk" and "taxonomy," and it refers to the on-the-fly classifications (called tags or keywords) that Internet users freely invent to categorize the objects with which they interact online. Social software makes these classifications available to other Internet users, often by means of a tag cloud, a list of user-developed tags. For this reason, folksonomy can be viewed as a distributed classification system.

Will our work be rendered obsolete? I'm not worried, but I am busy trying to keep up on the changes and the possibilities.

Slide 45 Contact information