Identification Needs in Developing, Documenting, and Indexing WSDOT Photographs

WA-RD 731.1

Barbara Endicott-Popovsky
Mike Simon

February 2010

Washington State Department of Transportation
Office of Research & Library Services

WSDOT Research Report

Research Report
WA-RD 731.1

Identification Needs in Developing, Documenting, and Indexing WSDOT Photographs

By

Barbara Endicott-Popovsky, PhD and Mike Simon
Center for Information Assurance and Cybersecurity
Information School
4311-11th Ave NE
University of Washington
Seattle, WA 98195

Jim Culp
Technical Monitor
Washington State Department of Transportation
Communications and Public Involvement Office

Prepared for:
Washington State Department of Transportation
Paula J. Hammond, Secretary

February 2010

TECHNICAL REPORT STANDARD TITLE PAGE

1. REPORT NO.	2. GOVERNMENT ACCESSION NO.	3. RECIPIENT'S CATALOG NO.
WA-RD 731.1		

4. TITLE AND SUBTITLE	5. REPORT DATE
Identification Needs in Developing, Documenting, and Indexing WSDOT Photographs	February 2010

7. AUTHOR(S)	8. PERFORMING ORGANIZATION REPORT NO.
Barbara Endicott-Popovsky and Michael Simon	

9. PERFORMING ORGANIZATION NAME AND ADDRESS	10. WORK UNIT NO.
Center for Information Assurance and Cybersecurity Information School; 4311-11th Ave NE; University of Washington Seattle, WA 98195	11. CONTRACT OR GRANT NO.

12. SPONSORING AGENCY NAME AND ADDRESS	13. TYPE OF REPORT AND PERIOD COVERED
Washington State Department of Transportation Olympia, Washington 98504 Project Manager: Kathy Lindquist, 360-705-7976	Research report
	14. SPONSORING AGENCY CODE

15. SUPPLEMENTARY NOTES

16. ABSTRACT

Over time, the Department of Transportation has accumulated image collections, which document important aspects of the transportation infrastructure in the Pacific Northwest, project status and construction details. These images range from paper photographs to extremely high resolution digital (or digitized) aerial photography, collected by many departments using film cameras, cell phone cameras, digital snapshot cameras, digital SLRs and special purpose geo-encoding stereographic cameras. Due to the diverse collection methods and technologies, as well as the growth of the use of imaging at the Department of Transportation – there are many separate archives of images with multiple incompatible access methods. This assessment identifies the current state of archiving and indexing images at the Department, summarizes a survey of key stakeholders in the current system, and recommends next steps toward developing an agency-wide image documentation and access system.

17. KEY WORDS	18. DISTRIBUTION STATEMENT
	No restrictions. This document is available to the public through the National Technical Information Service, Springfield, VA 22616

19. SECURITY CLASSIF. (of this report)	20. SECURITY CLASSIF. (of this page)	21. NO. OF PAGES	22. PRICE
None	None		

DISCLAIMER

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the data presented herein. The contents do not necessarily reflect the official views or policies of the Washington State Department of Transportation or the Federal Highway Administration. This report does not constitute a standard, specification, or regulation.

TABLE OF CONTENTS

EXECUTIVE SUMMARY

This report examines the current state of capturing, indexing, and archiving images and photographs at the Washington State Department of Transportation (WSDOT). The Technical Advisory Committee for the Indexing Photos Research Project (TAC) is interested in determining if any of the current practices adequately cover current, as well as anticipated, requirements. The TAC recognizes that the *status quo*—allowing multiple archives, multiple indexing technologies and multiple archive methodologies and policies—is inefficient, resulting in resource under-utilization. Exacerbating this situation is the pervasiveness of new technologies that allow any individual in the department to create and store images.

The recommendations presented in this report are based on 1) data collected from an online survey for which 126 responses were received (Appendix A), 2) interviews and emails with stakeholders using current systems, 3) meetings with the TAC, 4) analysis of external literature and 5) integration of the authors' experience with designing, creating and maintaining photo indexing systems.

Integrating these sources, the report concludes: 1) many informal methods exist for archiving images, most of which do not incorporate concepts of indexing, search, or an online interface, 2) some image silos do not directly lend themselves to re-organization within a single indexing structure, 3) although more than one interface has already been implemented for Stellent for which a site license exists, no single organizational plan guides its use, 4) regardless of what agency-wide solution is chosen, current owners of local archives should be encouraged to migrate existing photo silos to the solution of choice, thereby providing greater visibility for those images which

maintains their value; however, 5) if it is decided that some or all of the existing local archives be permitted to remain in their current state, then the agency should direct that any and all newly-created images be stored in the solution of choice.

Note: Migration to any solution, such as Stellent, should consider the development of consistent individual applications within that solution. This would allow for cross referencing as well as individual office security and usage. This is working well presently and the flexibility it affords is valuable.

INTRODUCTION

The Washington State Department of Transportation has an extensive collection of photographs, in both digital and print form, currently spread among various project offices that use different procedures for access and maintenance. An agency-wide, image-indexing system potentially could provide greater utilization of the collection; however, such a system would require considerable effort to create consistent metadata and a controlled vocabulary for describing and indexing images. Before investing in such an effort, a study of the existing WSDOT image collections was commissioned from a University of Washington team under the supervision of the WSDOT Technical Advisory Committee (TAC) (Appendix B).

This report summarizes that study and highlights issues to consider before developing a WSDOT image-indexing system. Information was compiled from an initial meeting with the TAC, individual discussions with key WSDOT staff, investigation of existing image collections, a survey of affected parties within the Agency, follow-on interviews with key stakeholders and an analysis of concepts and issues impacting indexing, organization, and retrieval of images. Preliminary project design recommendations are also provided.

REVIEW OF CURRENT AND PREVIOUS WORK

1.0 PURPOSE OF THIS STUDY

Prior to this study, a survey of other state DOTs was undertaken to gain insight into their indexing methods. The TAC study concluded that there appeared to be no consistent view among the states for how transportation departments organize and access their visual resources. Viewing this as an opportunity to pioneer an approach, the TAC decided to pursue development of their own internal system, commissioning this study to determine what existing WSDOT image collections use as approaches to indexing and retrieval and which, if any, might be extended to house the entire agency collection.

2.0 PROBLEM STATEMENT AND PREMISE

WSDOT currently has 8000 staff spread among headquarters and field offices across the State. Although coordination occurs with the Records Management Division to ensure proper maintenance of documentation subject to public disclosure or e-Discovery, procedures for organizing and accessing each collection are largely individualized at each location.

. Currently, millions of photographs and other visual materials are being maintained and accessed through a variety of formats and methods:

- Photographs/images in slides, 35mm film, digital and print formats
- Video, recordings
- Aerial photography
- As-built drawings, planned, and final maps
- Plans, real estate maps, geo-referencing data/docs for specific areas.

These visual resources exist across multiple applications and multiple servers, across multiple sites, and have no common indexing. In addition, different types of metadata may be collected for individual items, using different naming conventions and standards. Multiple titles, parcel numbers, etc., may be required for a single item. For many data attributes, there may be multiple description methods--for example, location may be described as city, or township, or latitude/longitude, or etc. In addition, certain types of data may change dynamically, creating a challenge to keep current.

The great diversity of how and where these collections are housed and maintained makes image visibility and access a challenge. Theoretically, the creation of a single WSDOT visual resources indexing system for all of these multiple archives could ease access and allow for greater utilization of the image collection.

3.0 REVIEW OF CURRENT PRACTICE

3.1 LITERATURE REVIEW

The broad adoption of multiple forms of digital image capture—digital cameras, scanners, and print-to-image applications—has created a challenge for electronic archiving and indexing. Images are large, in comparison to text files, with some high resolution images being several gigabytes—the equivalent of 600,000 text pages.

The size of these images is problematic for indexing and database systems which were designed to handle text documents and typically embed such documents into the database corpus itself. Text-based systems do not function well if individual elements are equivalent to hundreds of thousands of pages of text (Silberschatz, 1990).

Computers are not yet able to efficiently process the information content of images in a way that lends itself to meaningful text-based search and retrieval (Chiueh, 1994). Fortunately, most modern image formats incorporate metadata into the structure of the image file itself. Linking external metadata from a database to each image allows for efficient indexing and retrieval in a wide variety of systems and methodologies (Ogle, 1995 and Grosky, 1994).

Metadata is useful for describing the contents and context of a file without actually changing the file's contents. While work has been done on the use of metadata for specific non-textual files such as music, images, and video, there is utility for metadata that transcends all file types and disciplines (Weibel, 1998).

Metadata standards discussed in the Dublin Core papers (Weibel, 1998), as well as the Visual Resources Association Core, REACH (Record Export for Art and Cultural Heritage), and EAD (Encoded Archival Description) metadata schemas, all support the organization and description of image data. While domain ontology schemas like VRA and REACH work well in this context, foundation ontologies like Dublin Core and EAD are also capable of supporting image organization and description (Greenberg, 2001 and Hruby, 2005). The literature suggests that if a foundation ontology (Kent, 2001) is already in use elsewhere in an organization, it could be used successfully in image indexing.

At the enterprise level, indexing, search, and retrieval functions are viewed as the purview of database applications. There is a general recognition that organizations have difficulty proactively developing metadata, leading to passive indexing and the use of data mining and other information discovery processes. These strategies do not work well

for images because they depend on what information can be discovered about an object from an analysis of its context (date, time, size in pixels, server location) and information content, which is difficult to characterize more specifically than generalizations about color, shapes and guesses about content (Lehmann, 2005).

More recently, the plethora of medical images and the need to store, retrieve, and share these images across multiple user communities has advanced the state of image retrieval to include concept-based retrieval methods. One example would be indexing images using concepts extracted from the associated captions; however, this is extremely laborious to perform manually. Another is an automated technique to map the unstructured ("free") text of figure captions to concepts in a set of controlled vocabularies. Methods such as these can enable the radiology community to access more effectively the vast amounts of radiological image data being published online (Kahn, 2009).

3.2 SUMMARY OF EXISTING COLLECTIONS

Image collections exist in many forms, across multiple departments within the Washington State Department of Transportation. Two primary WSDOT collections have advanced indexing tools that could be a starting point for developing an agency-wide, image-indexing system.

3.2.1 Primary WSDOT Collections

3.2.1.1 Aerial Photography Collection

(http://www.wsdot.wa.gov/MapsData/Aerial/intro.htm)

The Aerial Photography image archive contains over 750,000 images of the Pacific Northwest dating from 1933 to the present in B/W, true color and color

infrared film. These images are not available online and must be ordered through the Aerial Photography division, which is a profit-based division within the Agency. The collection uses multiple indexing structures and metadata schemas.

3.2.1.2 Stellent Image System within the Records Management Division

Stellent is an internal imaging system containing over 2 million images, including a large collection of historical images. Each project within Stellent has its own indexing structure specific to the customers' needs. Photos can be retrieved from the system, but the system does not allow searching across projects. Where there are contract numbers, such as with "As Built" drawings or "Right of Way" maps, there are references within the photo application that list the contract number if there is one related to the photo. The system recognizes two types of media – physical objects versus electronic.

The following collections have less developed indexing structures and are categorized as either internally housed within WSDOT divisions and projects, or externally housed through the State including the Secretary of State Archives, the University of Washington, or commercial sites such as Flickr.com. The creation of any visual resources system should consider incorporating or referencing these external collections, including their associated metadata.

3.2.2 Internally-housed Resources: Individual Project Photo Galleries

Many individual projects have their own photo gallery available on the WSDOT website:

3.2.2.1 Tacoma Narrows Bridge Tolling

(http://www.wsdot.wa.gov/Operations/Tolling/TNBTolling/photogallery.htm)

3.2.2.2 SR 520 - Bridge Replacement and HOV Project - Photo Galleries

(http://www.wsdot.wa.gov/Projects/SR520Bridge/Photos/newgallery.htm)

3.2.2.3 Alaskan Way Viaduct and Seawall Replacement Program - Photo Gallery (http://www.wsdot.wa.gov/Projects/Viaduct/Gallery.htm)

3.2.2.4 SR 104 - Hood Canal Bridge - Photo Gallery (http://www.wsdot.wa.gov/Projects/SR104HoodCanalBridgeEast/photogallery.htm)

3.2.3 Externally-housed Resources

3.2.3.1 Pavement Division photos at UW Transportation Media Library (http://photoview.ce.washington.edu/)

The UW Transportation Media Library consists of a web-based collection of transportation images. The WSDOT Pavement Division may be using the site to catalog images within their department. The collection also has a list of keywords that may be useful in the development of vocabularies for indexing WSDOT images.

3.2.3.2 WSDOT Flickr photostream (http://www.flickr.com/photos/wsdot/)

The WSDOT Flickr photostream is publicly available and contains 2000+ items, 82 sets of images, and more than 150 tags. The photostream is accessible through a WSDOT blog: http://wsdotblog.blogspot.com/.

3.2.3.3 WSDOT YouTube Videos (http://www.youtube.com/user/wsdot)

WSDOT has 31+ videos hosted on YouTube, which are also linked from the WSDOT blog.

3.2.3.4 Washington State Archives Image Collection

(http://www.digitalarchives.wa.gov/RecordSeriesInfo.aspx?rsid=22)

While no current WSDOT photos may be located in this collection, it is a resource for historic photos.

3.2.3.5 UW Libraries: Historic WSDOT Photos

(http://content.lib.washington.edu/index.html)

The number of photos within the UW Libraries image collection is unclear. Also unclear are rights associated with those photos, or whether any agreements exist between WSDOT and UW for the use of those photos.

A survey of existing collections will be useful in understanding, quantitatively, how use breaks down into categories.

4.0 OTHER RESOURCES AVAILABLE

WSDOT has a number of data management resources that should be reviewed before development of the final image library. These resources consist of public metadata sets for documents, content management systems, the WSDOT MS SharePoint implementation, and WSDOT's data catalog. These resources contain standardized datasets and vocabularies that could be useful for any future agency-wide image catalog.

RESEARCH APPROACH/PROCEDURES

The objectives of the WSDOT Imaging Assessment Project are 1) to identify WSDOT user requirements for photo indexing, storage and retrieval and 2) to analyze existing WSDOT indexing and storage systems to determine if any meet enterprise requirements. If the project is unable to identify an existing WSDOT software system capable of meeting the entire needs of the enterprise, a second project may be required to identify such a system.

As indicated earlier, WSDOT has an extensive collection of images representing all aspects of the State's transportation infrastructure. These collections are not easy to find and access. The lack of standardized organization results in substantial search efforts and increased burden on server space due to unnecessary duplication.

This study is an initial step toward an online repository that could offer ease of access and use.

1.0 METHODOLOGY

To discover what methods, tools and procedures are currently used to archive, index and retrieve image information, a large cross section of the WSDOT user community was asked to participate in an online survey (Appendix A), which gathered details regarding use of images in daily work, the systems used to access those images, and what procedures were used to capture image information and associated metadata. The survey was completed in November, 2008, with 126 total respondents from the 447 employees approached.

Additionally, two interviews were conducted with the TAC to verify survey information gathered and to further understand the committee's perspective as active users of image data.

In order to determine what functionality a future system should include and how it should be organized, it was important to identify how WSDOT currently uses visual resources. The following were revealed through the techniques described above:

1. Public outreach via websites, factsheets, and reports. Many of these are produced through the Communications Office, in coordination with the department responsible for the project of interest. Requests for images also originate from outside the organization, often through the WSDOT Library.

2. eDiscovery which can consist of requests from multiple departments for any, or all, items related to a particular project. Currently, the office responsible for eDiscovery is the Records and Information Services Division, which uses the Stellent system. Providing efficient access would reduce response costs related to these requests.

3. Project development and maintenance. Projects require easy access to current and historic maps, project designs, and site images. Many historic images have already been digitized and often date back to the early part of the last century; however, many others have not been digitized and are more difficult to acquire.

A survey of utilization was helpful in understanding, quantitatively, how use breaks down into the above categories.

2.0 SURVEY

The survey, consisting of 95 questions, was conducted online using a tool hosted at the University of Washington. The TAC assisted with writing the survey and getting sufficient participation.

The approach taken to the development of the survey assured that the survey:

- Had clarity of purpose.
- Had a narrow scope.
- Included incentives for survey completion – e.g., public disclosure motivation.
- Identified supporters within WSDOT who could support completion.
- Made clear that WSDOT was not trying to beg an outcome by naming things in ways that are not being used.
- Did not impose someone else's agenda.
- Emphasized the need to assess the extent of the problem.
- Did not emphasize identification of a solution.
- Focused on identifying 'current best practices' in order to learn from them.
- Assumed participants had no technical skills, knowledge.
- Kept terminology simple and straightforward.

The survey tool was designed to collect the followings pieces of information:

- Sources used for acquiring WSDOT photos.
- Parameters needed for searching photos.

  e.g. location, date, keyword, dept., etc.

- Current means and authorities for accessing photos,

  e.g. repository on desktop.
- ORG code – to identify differences in responses per division.
- Use of other, external systems.

(Note: There are no agreements in place for accessing such images, like those residing at UW.)

A draft survey was reviewed by key members of the TAC, delivered through a UW Catalyst WebQ: (https://catalysttools.washington.edu/webq/survey/nlou/57966), and opened to the public for response. The entire project team had full administrator access to the survey. The TAC assisted with selecting key staff to distribute the survey, which included representatives from key departments such as the Office of Information Technology, the WSDOT Library, Communications, Records Management, and selected specific projects. The final list of survey questions is found in Appendix A.

The following survey protocol and timeline were followed:

A. Survey Preparation

   1) Initial review by TAC/WSDOT prior to distribution

   2) TAC/WSDOT-compiled initial distribution list

   3) Survey created in UW Catalyst WebQ

B. Survey Administration

   1) Survey distributed to key staff via TAC

   2) 2-3 weeks allowed for responses

C. Results Compiled and Analyzed

   1) Responses compiled, organized, and summarized

2) Participants identified for in-depth interviews

3) Interview questions generated based on survey responses

D. In-depth interviews conducted (see below)

E. Findings summarized in report

Note that the online survey was purely voluntary, therefore some bias can be assumed, based on self selection by those interested enough in photo indexing, and involved enough in photographic projects, to invest time completing the survey. The Advisory Committee invited 447 staff to participate. One hundred and twenty-six responses were collected.

3.0 INTERVIEWS

Interviews provided insight into advanced practices being used within WSDOT, detailed information on how images are currently organized and accessed in WSDOT, identified key image handlers and assessed their processes for handling images from acquisition to indexing to publishing, and gave a fuller understanding of the environment in which these systems operate.

A Phase II follow-on project could gather additional information to include:

- Nomenclature, naming conventions, vocabularies in use (e.g. Flickr tags, Transportation Research Thesaurus).
- Identification of a document management lifecycle, including the phase of an element during the lifecycle process or whether it is has a public vs. internal use.
- How originals are identified.

- How sensitivities are indexed, i.e., TRT with 10K preferred terms – which lacks operational insight due to few lead-in terms.

FINDINGS/DISCUSSION

Survey results, literature reviews and discussions with stakeholders at the WSDOT all suggest that images used by the Agency are an important resource which may be underutilized, or in some cases may even lose nearly all its value, unless a systematic approach is taken to provide interfaces which are uniform throughout the organization. This uniform system should be backed by policies and guidance for acquisition, archiving, and use of photographs. The organizational component of an initiative to create such a uniform system is even more critical than the selection of any one particular technology. It is the unifying policy and operational guidelines that will create and maintain a single view of the image corpus, which will be assisted by the chosen technology, not driven by it.

Interviews with stakeholders and the TAC reflect the information collected in the online survey, adding detail regarding some of the more sophisticated tools and Agency use of photographs. Currently, the Records Management and Aerial (cost-center) divisions appear to have the most developed systems.

These interviews reflected a growing unease among stakeholders regarding the lack of uniform guidelines and operational recommendations in the area of photographic archiving and indexing, emphasizing the concern that significant work product is potentially at risk of losing value to the organization if photographs are maintained in individual silos without universal indexing and access. While interviews are subjective and considered only as anecdotal evidence, these observations closely align with the reasons for requesting this study.

1.0 SURVEY RESULTS

The following analysis describes the Catalyst Survey results:

To maintain the anonymity of respondents, a sanitized version of the responses is being made available with this report (names, email addresses removed).

Questions 1-5 (General Information):

One hundred and twenty-six individuals responded to the survey. The survey results accompanying this report provide an overview of the diversity of departments and job roles of those who responded. As mentioned earlier, names have been withheld to preserve anonymity.

Questions 6-12 (Current Image Collections):

The diversity of responses to questions about use of images, size of collections, formats, maintenance role, protocols used, make characterization/summary difficult. It did lead to the conclusion that it will be difficult to move all of these collections to a single repository and indeed it may not be cost effective to do so. Instead, incentivizing users, rather than levying a standard, is the likely better approach. Incentivizing would include demonstration of efficiencies, cost-saving opportunities, etc.

The most interesting results were derived from Question 11b about attributes relied upon for retrieval. Among those who responded, the breakdown of the attributes they use to identify photos follows:

- Date taken 75
- Location 64
- Image name 52
- Project 48

- Keywords 25
- Other 17
- Credit, Version - negligible

When describing the location of a resource, the responses broke down according to the following list. What is most obvious about this data is that there is a lot of variation in how location is described, with "other" being the largest category.

- Other 57
- SR-Milepost 51
- Street address 17
- Lat/long or X-Y Coordinate 7

Questions 13-21 (Current Use of Images)

Use, frequency, source (i.e., external vs. internal), access were queried and the results can be found in the accompanying list of survey responses. Respondents used photos primarily for reports and presentations. Responses indicated the following distribution of uses:

- Reports 87
- Presentations 78
- Others 50
- Web site 44
- Non-litigious public requests 28
- Factsheets/marketing materials 23

Common criteria for retrieving photos were distributed widely as indicated below: The main attributes relied on were SR-Milepost, date, location, title, project title.

- SR – Milepost 54
- Date 53
- Location 50
- Title 42
- Project Title 41
- Contract Number 30
- Keywords 29
- Other 15
- Bridge Number 13
- Program/Division/Office or Org Code 10

- Lat/Long 6

The most interesting question for our purposes was Question 18 which asks responders to rank potential features of a new system. These were ranked as follows.

Responders ranked each attribute by scoring to the following standards:

1= Not necessary
2= Not essential, but nice to have
3= Essential for an image catalogue

VOTES / FEATURE

307	Ability to search for images by topic/keyword (like using Google or Yahoo)
305	Ability to view and download high resolution images
306	Ability to browse for images by category
292	Ability to perform guided searches, selecting from known keywords
281	Viewable on-line through a web interface
270	The catalog should allow batch uploads of images to expedite the uploading process
266	Viewable to all WSDOT staff
256	Viewable to the public
254	Allow photo maintenance staff to perform basic editing functions for images (e.g. resizing, cropping, color modification, etc.).
249	Maintained (e.g. images uploaded, edited) on-line through a web interface.
247	Allow photo maintenance staff to add their own keywords to images.
244	Include a pre-existing list of keywords (e.g. Transportation Research Thesaurus) for staff to use in describing individual images.
227	The catalog should also track images in non-digital form (e.g. slides, photographs).
189	Allow all WSDOT staff to add their own keywords/tags to images that they did not upload.
172	The catalog should only track digital images
123	Allow external (public) users to add their own keywords/tags to images.

The following descriptive elements, along with the number of respondents, were identified in Question 19 as always being necessary for tracking images:

- Date taken 115
- Location 105
- Description 80
- Image name 78
- Related Project 67
- Keywords or tags 47
- Credit 16
- Latitude/longitude 14
- Other 11

Questions 21 (Open-ended response)

Again, readers are referred to the attached survey results to get a sense of the open ended responses to the final question.

2.0 NEXT STEPS

Ultimately, WSDOT wishes to develop a system that helps WSDOT employees find visual resources in an efficient and accurate way. **Findability** is simply the quality of something being locatable and observable at two levels. At the item level, something is "findable" if you can easily locate it. At the system level, something is "findable" as a result of ease of navigation and retrievability within a particular system. Within a visual resource system, findability can be enhanced through a common indexing system that uses consistent metadata standards.

The following section describes some of the requirements to consider for developing a common indexing system.

2.1 SYSTEM DEVELOPMENT

Preliminary discussions with WSDOT staff have indicated that due to visual resources existing across many offices and many systems, it appears that having a common indexing structure is possible, but one common repository is not likely. An indexing system would merely need to define commonalities for description, potentially centered on common metadata structures described below. For the first phase of a WSDOT Image Library of the Future (WILF) project, a complete inventory of existing collections may not be necessary. Instead, a cross section of what exists may be sufficient to create an initial indexing system. While a common repository could be listed as an

alternative solution, a gap study should be performed before initiating such a project to help identify needs, existing sources, gaps, and alternatives.

2.2 INDEXING PROCEDURES

If an indexing system is the best solution, rather than a single repository, then a project to define indexing procedures and standards would be the best next step. Indexing is the process used to identify, describe, and label the contents of a document or image so that it can be retrieved later, during search. It is used to build metadata. What protocol to use for indexing images, specifically what elements are required to describe images submitted to the library, should be explicitly defined in the metadata.

For the WILF project, an individual should be identified who will be responsible for indexing images, based on the required metadata. The indexer may be an experienced professional librarian, familiar with using vocabulary tools, or a WSDOT staff member, with expertise in specific projects. Past research has found little difference whether a subject or non-subject specialist is assigned to the application of terms. (Most variation is due to poor instruction on how terms should be applied, different perceptions among indexers, and changes in perception over time.)

Depending on which organizational method is used (e.g. thesaurus, tags), different degrees of training may be necessary for the indexer to fully understand how to effectively and correctly use the selected vocabulary. This is essential to ensure adequate retrieval of images during the search process.

2.3 VOCABULARY TOOLS

Descriptors or terms are assigned during indexing from a number of vocabulary tools based on guidelines for the selection and application of appropriate terms. The

following vocabulary tools exist for generating descriptive terms and should be considered for the WILF project (Jörgensen, 2003).

2.3.1 Thesaurus

The purpose of a thesaurus is to facilitate the indexing and searching of documents. A thesaurus should contain all of the terms needed to describe a given document that falls within the subject domain of the thesaurus. These terms are known as "preferred terms."

A thesaurus should also show relationships between preferred terms, such as broader, narrower, or related terms, and identify synonymous terms, pointing the user to the preferred term that is used in the thesaurus.

The use of a thesaurus is instrumental in indexing, providing indexers with a controlled vocabulary of terms to apply. This allows consistency in terms, and reduces the potential for variation that may result if indexers were to create their own terms. Documentation within the thesaurus, such as scope notes, provides the indexer with an understanding of how a term should be applied during the indexing process. This also allows for consistency among all indexers as they apply the same set of terms to different documents.

Users employ a thesaurus in a number of ways. First, they may actively review the content and structure of a thesaurus to acquire an understanding of how documents within a database are indexed. This allows identification of relevant search terms, and determines broader, narrower, or related terms. Users may also employ a thesaurus when reviewing results from a particular search.

Within the field of transportation, the Transportation Research Board has created the *Transportation Research Thesaurus*, which could be used as a potential resource for the WILF project; however, the thesaurus is limited in its use of lead-in terms and may not represent the best option due to its limited vocabulary. The thesaurus can be found at http://trt.trb.org/.

2.3.2 Subject Headings

Subject headings provide a set of controlled terms for subject access. They may be organized using shallow hierarchies, with varying consistencies, depth, and breadth within and across different systems. Unlike thesauri, they may not indicate relationships among terms as extensively.

Most libraries make use of the Library of Congress Subject Headings (LCSH) and the Anglo-American Cataloging Rules 2 (AACR2) to consistently apply subject headings for graphic materials; however, this approach is often problematic for images due to the depth of indexing and different levels of access.

2.3.3 Authority Files

Authority files help indexers and searchers determine what variant names exist for an entity, including personal, organizational, and geographic names. Simply put, they are pre-determined picklists for specific data fields. For example, an authority file for WSDOT departments may consist of all department names and abbreviations used within WSDOT. Indexers would then select the appropriate department from this list.

For WSDOT, authority files may already exist for organizational names, and geographic units and locations. These files may originate in other data management

systems within the Agency. It would be important to assess the importance of these files for use within the metadata standards developed for the WSDOT image library.

2.3.4 User-generated Tags or Keywords

Tags and keywords (also known as folksonomies) are single words or phrases assigned to a particular item. Tagging is considered a flexible and easy-to-use system of organization because users themselves generate tags, so they can use their own preferred terms to identify an item. While tags are user-friendly, they do create problems in semantics, as well as findability.

People may also use different words to describe the same thing. When different words are used, or different meanings are implied by the same word, it makes it difficult during the retrieval process to find all items associated with a particular tag. This results in some items being irrelevant, due to different meanings for the same word, while other items may be missing, because they have been indexed with a different tag.

WSDOT's existing photo collection within Flickr is an example of an indexing system that uses user-generated keywords to apply to images. A complete list is available at: http://www.flickr.com/photos/wsdot/alltags/. A close examination reveals one of the drawbacks of user-generated keywords. In the keyword list, Mt. Rainier is identified in two ways: "mtrainier" and "mountrainier." This is a common issue in user-generated keyword lists and creates a problem during retrieval, as you would have to use both keywords to see all images associated with Mt. Rainier.

2.4 METADATA STANDARDS

The library and information sciences have produced a number of indexing guidelines (e.g. the Anglo-American Cataloging Rules II) and data structure standards

(e.g. MARC format). While these are useful for bibliographic records, their utility in other types of systems may be less appropriate.

The following list of metadata standards summarizes options for consideration in the development of a visual resources indexing system. Each standard is officially endorsed by the Metadata Encoding Transmission Standard Editorial Board, confirming its compliance with nationally-recognized metadata standards. For a more detailed overview of metadata, see the online guide maintained by Getty Images (http://www.getty.edu/research/conducting_research/standards/ intrometadata/).

2.4.1 VRA Core Version 4.0

(http://www.vraweb.org/projects/vracore4/)

The Visual Resources Association (VRA) has been working to create standards to describe images since the 1980's. The VRA Core 4.0 was published in 2007. It provides a set of recommended metadata elements, as well as suggestions for how elements can be hierarchically structured.

2.4.2 Dublin Core

(http://uk.dublincore.org/schemas/xmls/)

Version 1.1 of the Dublin Core Metadata Element Set identifies fifteen elements for describing a wide range of resources. The purpose of the limited element set is to support cross-disciplinary resource discovery. The fifteen elements include: *Contributor, Coverage, Creator, Date, Description, Format, Identifier, Language, Publisher, Relation, Rights, Source, Subject, Title, and Type.*

2.4.3 Resource Description Framework

(http://www.w3.org/RDF/)

RDF is a specification from the World Wide Web Consortium (W3C) to describe resources on the web, based on their properties, using XML syntax. It is a framework for resource description that uses RDF schemas based on XML to identify resources and describe the relationships among them. RDF schemas are created based on the needs of a community of application. Schemas can also incorporate other metadata standards such as Dublin Core in the specific description of different elements.

While there may be many methods currently in use, a major requirement is that the system support associated metadata that meet the Visual Resources Association Core data standards, currently at version 4.0. The VRA defines a metadata structure to be associated with images that includes an XML standardized structure to support interoperability and interchange

2.5 Extensibility

Extensibility is the ability of a system to anticipate and prepare for future growth. This could be achieved through adding functionality, accommodating additional formats, and adapting to future needs. For a visual resources indexing system within WSDOT, some areas of growth to consider in extensibility planning may include the following:

- Incorporation of new image management systems within and beyond WSDOT
- Inclusion of additional formats other than those initially defined
- Expanding access to a broader user base
- Communications with external systems.

2.6 Usability

Usability can be defined as the ease with which a person can use a particular tool to achieve a certain goal. As mentioned above, WSDOT has a number of uses for visual resources including outreach and communications, eDiscovery, and project maintenance and development. Within each of these uses, there are specific ways in which resources need to be searched and accessed. A future system will need to anticipate these needs and design functionality around how to make image access more efficient in order to increase the utility of the system to meet department goals.

2.7 Organizational Issues

The implementation of the WILF will have a number of institutional impacts as WSDOT staff learn to use the system, as both content providers and users. Some questions to consider as recommendations are being developed:

- What will be the role of WSDOT departments and individual staff?
- How can imaging standards be institutionalized across the Agency?
- What are current copyright restrictions; how should the system handle them?
- Is versioning or phasing of resources an issue; how can related rules be built into the system?

The Washington State Library Digital Images Initiative has assembled a website summarizing best practices for developing digital collections. The website highlights project management, collection development, and technology issues and is an important resource. The guide is available at: http://digitalwa.statelib.wa.gov/newsite/best.htm

CONCLUSIONS

A summary of conclusions drawn from this study are as follows:

1. There are many informal methods in use at WSDOT for archiving images, most of which do not incorporate concepts of indexing, search or in most cases even interface (e.g. putting images on a network shared drive with no outwardly logical organization.)

2. There are image silos that do not directly lend themselves to re-organization within a single indexing structure, or are already part of a mature indexing and retrieval structure (e.g. Flickr postings and Aerial Imaging, respectively.)

3. There are at least two firmly held, divergent views of what should be done to resolve the image accessibility issue. Both views are valid—one sees the problem from the library/customer service viewpoint; the other sees the problem from the project management viewpoint. These views may be difficult, or even unnecessary, to reconcile.

4. Striving for a single indexing system might be a worthy goal so that the Agency converges on common terms, making retrieval easier, but having a single technical environment may not meet all of the needs of the different units in the Agency and it may not be cost effective.

5. Providing support to the general public in the volumes and frequencies needed is better supported through environments like Flickr, where millions of images are requested each year. The University of Washington image retrieval system is worth studying further in this regard.

6. An Agency site license exists for Stellent, an image archiving system with all the capabilities described in survey responses and interviews. More than one interface has already been implemented for Stellent, although no single organizational plan so far guides its use.

7. While one technical environment may not meet all of the needs of the Agency, the number of different solutions could be reduced to simplify retrieval.

8. Organizational issues are as important in resolving the problem as technical solutions. It will require effective leadership to move the Agency from the plethora of systems now in existence to a more manageable number. Well-written and coordinated policies and extensive collaboration will be essential.

RECOMMENDATIONS

While this study does not presume to recommend a single solution, there are practical answers that might be implemented efficiently without necessitating full Phase II development efforts to create a new, enterprise-wide solution.

Given these economically constrained times, we recommend that the WSDOT consider the following:

1. Develop a single indexing system/taxonomy for use in describing images used by the Agency.

2. Develop technology independent policies, procedures and guidelines for the collection and archiving of images with some attention to required metadata and optional metadata to be associated with each image.

3. Develop a plan to implement further adoption of the already-owned Stellent infrastructure to accommodate as many of the image users as possible, taking advantage of the already existing Agency-wide license. There should be consistent guidance provided across the agency on how to use Stellent.

NOTE: The original Oracle Stellent license was limited in scope, and over the course of this study was expanded to cover use by the entire organization for reasons not directly related to this project. While there are a number of competing products on the market like Interwoven Metatagger, FileNet and Open Text , Stellent consistently scores well in side-by-side tests with these other products, and has the advantage of incurring no additional purchase costs to the Agency, as well as taking advantage of in-house expertise.

4. Launch a study into the University of Washington image retrieval system to identify and adopt practices that will improve retrieval from archives such as Flickr, now being used for customer service. The goal is to make public access easier.

5. Determine whether the plethora of existing silos of images outside of these two main environments should be imported, left as-is or transitioned, on-going, to the new approaches just described. In this analysis, exercise a bias toward reducing and simplifying the numbers and types of archives currently in existence in order to make retrieval easier.

6. Regardless of what technical solution/s is/are chosen, owners of current archives should be encouraged by policy to migrate existing photo silos into any newly named standard system, thereby providing greater visibility of current archives, maintaining their value.

7. If some archives are left in their current state, then policy should direct that any new images created should populate the tool/s designated as standards in the Agency.

8. Continue to support and nurture the TAC committee as the collaborative entity that will develop and promote Agency-wide policy that will lead to a less chaotic environment for image retrieval.

These tasks are substantial undertakings, but potentially less disruptive and likely less expensive than adopting a new platform entirely. We do not recommend maintaining the *status quo*. We expect that will lead to useful images being lost due to a lack of a global indexing and retrieval systems, devaluing the collection.

ACKNOWLEDGEMENTS

The authors would like to express their appreciation for the time and effort WSDOT employees spent responding to surveys, attending committee meetings and responding to email Q and A by all of the involved WSDOT staff. Your thoughtful comments and explanations made the information gathering part of this report a pleasure and improved the quality of our observations immeasurably. We would like to extend our very special thanks to Kathy Lindquist for coordinating, managing, and generally keeping the process going on the WSDOT side.

REFERENCES

Chiueh, Tzi-cker, 1994. "Content-Based Image Indexing" Proceedings of the 20th VLDB Conference, King Fahd University of Petroleum and Minerals

Gill, T., Gilliland, A.J., and Woodley, M.S. (). Introduction to Metadata: Pathways to Digital Information. Online Edition, Version 2.1 Available2.1 Available: http://www.getty.edu/research/conducting_research/standards/intrometadata/index.html

Greenberg, Jane, 2001. "A quantitative categorical analysis of metadata elements in image-applicable metadata schemas," Journal of the American Society for Information Science and Technology Volume 52 Issue 11, pp. 917-924

Grosky, William I. and Farshad Fotouhi and Ishwar K. Sethi and Bogdan Capatina, "Using metadata for the intelligent browsing of structured media objects," ACM SIGMOD Record archive Volume 23, Issue 4, December 1994, pp. 49-56

Hruby, Pavel "Ontology-Based Domain-Driven Design," Proceedings of the 2005 Object-Oriented Programming, Systems, Languages and Applications (OOPSLA) Conference, San Diego, CA

Jörgensen, C. Image Retrieval. Scarecrow Press, Inc. Lanham, MD: 2003.

Kahn, Charles E. and Daniel L. Rubin "Automated Semantic Indexing of Figure Captions to Improve Radiology Image Retrieval," Journal of the American Medical Informatics Association, 2009, Volume 16, pp. 380-386 Available: http://jamia.bmj.com/content/16/3/380.full

Kent, Robert E. IFF Foundation Ontology, SUOWG IEEE 2001. Available: http://suo.ieee.org/IFF-Foundation-Ontology.pdf

Lehmann, Thomas M and Güld Mark O; Deselaers Thomas; Keysers Daniel; Schubert Henning; Spitzer Klaus; Ney Hermann; Wein Berthold B, 2005. "Automatic categorization of medical images for content-based retrieval and data mining," Computerized medical imaging and graphics Volume 29, pp, pp 143-55

Ogle, V.E. and M. Stonebreaker, 1995, "Chabot: retrieval from a relational database of images," IEEE Computer, Volume: 28, Issue, Issue: 9 September 1995 pp. 40-48

Silberschatz, Ali and Michael Stonebraker and Jeffrey D. Ullman, 1990. "Database systems: achievements and opportunities," ACM SIGMOD Record archive Volume 19 , Issue 4, December 1990, pp. 6-22

Washington State Library, Digital Images Initiative. Digital Best Practices. Available online: http://digitalwa.statelib.wa.gov/newsite/best.htm

Weibel, S., J. Kunze, C. Lagoze, M. Wolf, 1998 "RFC2413 Dublin Core Metadata for Resource Discovery" RFC Editor, USA

APPENDIX A: SURVEY DESIGN

WSDOT Image Library
Questionnaire of Current and Future WSDOT Image Maintenance Procedures

The purpose of this questionnaire is to assess current procedures used to maintain images related to WSDOT activities and programs, and identify potential uses of a future image catalog. The questionnaire consists of four sections with a total of 21 questions. The questionnaire should take about 10-15 minutes to complete.

We will be using the results to propose design recommendations for a future WSDOT image catalog. If you have any questions about the project, please contact Barbara Endicott-Popovsky. Thank you very much for your time!

General Information

1. Name [short answer]

2. ORG Code [short answer]

3. Position Title [short answer]

4. Please describe your primary job responsibilities, listing up to ten tasks

    - [short answer]
    - [short answer]
    - [short answer]
    - [short answer]
    - [short answer]
    - [short answer]
    - [short answer]
    - [short answer]
    - [short answer]
    - [short answer]

5. We may choose to conduct additional phone interviews with certain respondents to gain a deeper understanding of current procedures used to collect, store, and access visual resources in WSDOT.

   Would you be willing to participate in these additional phone interviews? [yes/no]

   If so, please provide your e-mail address so we can schedule a time [short answer]

Current Image Collection

6-7. Do you currently have access to a collection(s) of visual resources related to WSDOT work? [yes/no]

- Name of the collection [short answer]

- Description of the collection [short answer]

- Is this collection available electronically? [short answer]

- Location or URL (if applicable) [short answer]

- Size of collection (e.g. number of individual images, photos, videos, etc.) [short answer]

- Do you have another collection to add? [short answer]

- Name of the collection [short answer]

- Description of the collection [short answer]

- Is this collection available electronically? [short answer]

- Location or URL (if applicable) [short answer]

- Size of collection (e.g. number of individual images, photos, videos, etc.) [short answer]

- Do you have another collection to add? [short answer]

- Name of the collection [short answer]

- Description of the collection [short answer]

- Is this collection available electronically? [short answer]

- Location or URL (if applicable) [short answer]

- Size of collection (e.g. number of individual images, photos, videos, etc.) [short answer]

8. Please describe what format these images are in, and what percent of your collection they represent? [matrix of formats/percentages]

- Digital Images
- Printed photographs
- Negative or slides
- Film and video
- Other formats (indicate format at percentage)

9 Describe how you access non-digital resources (e.g. photographs, slides, video)? [short answer]

10. Do you currently have staff dedicated to photo or visual resources management and/or maintenance? [short desc]

The following questions pertain to those who organize and maintain their own photos:

11a. Do you, or your staff, add any descriptive information to the images you maintain, either in digital or hard copy form? [long desc]

11b. Which of the following attributes do you include: (select all that apply)

- Location
- Key words
- Date Taken
- Image name
- Project
- Credit
- Other (please list)

11c. How do you describe the location of a resource? (select all that apply)

- SR Milepost
- Street Address
- Lat/long or X-Y Coordinate
- Other (please list)

12a. Are there existing WSDOT or professional rules or protocols you use to maintain or index your image collection? These may consist of rules regarding metadata, attributes, or naming conventions. [short answer]

12b. What specific protocols do you use? Please provide a link if they are available on-line. [short answer]

Current Use of Images

13. How do you use the images in your collection? Please review this list carefully and help us identify other potential uses that we may not have considered. (select all that apply)

    - Presentations
    - Reports
    - Factsheets/marketing materials
    - Web site
    - Non-litigious public requests
    - Other (please list)

14. How often do you use images in your collection for the purposes above? [short answer]

15. Are there any existing image collections you use when looking for an image? If so, please provide their name and URLs (if applicable). These collections may be internal or external WSDOT resources. [short answer]

16. What are some of the common criteria you use to find an image? (select all that apply)

    - Keywords
    - Location
    - Date
    - Project title
    - SR – Milepost
    - Lat/Long
    - Bridge number
    - Contract number
    - Title
    - Program/Division/Office or Org Code
    - Other (please list)

17. Who do you contact if you don't have, or cannot find what you need? [short answer]

Potential Use of Images

18. Evaluate the following statement that describes potential features of an image catalog. Rank these statements in terms of how important they are to you according to the following scale:

    Ranking scale:
    1 = Not necessary

2 = Not essential but would be nice to have
3 = Essential for an image catalog

A digital image catalog should include the following features:

- Viewable on-line through a web interface
- Viewable to all WSDOT staff
- Viewable to the public
- Ability to view and download high resolution images
- Ability to search for images by topic/keyword
- Ability to browse for images by category
- Maintained (e.g. images uploaded, edited) on-line through a web interface.
- Include a pre-existing list of keywords (e.g. Transportation Thesaurus) for staff to use in describing individual images.
- Allow photo maintenance staff to add their own keywords to images.
- Allow photo maintenance staff to perform basic editing functions for images (e.g. resizing, cropping, color modification, etc.).
- Allow all WSDOT staff to add their own keywords/tags to images that they did not upload.
- Allow external users to add their own keywords/tags to images.
- The catalog should also track images in non-digital form (e.g. slides, photographs).
- The catalog should only track digital images
- The catalog should allow batch uploads of images to expedite the uploading process.

19. The following descriptive elements should always be tracked for images: (select all that apply)

    - Image name
    - Date taken
    - Location
    - Related Project (if applicable)
    - Description
    - Latitude/longitude
    - Credit
    - Keywords or tags
    - Other? Please list _____

20. Have you come across other image libraries or catalogs that have desirable qualities that you would like to see in a WSDOT catalog? If so, please describe the name, source, and provide a link to the catalog, if available. [long desc]

21. Is there anything else you would like to share about your use of images and visual sources within WSDOT? [long desc]

Thank you very much for taking the time to complete this survey. If you indicated that you would be willing to participate in additional phone interviews, we will contact you shortly to schedule a time.

APPENDIX B:

WSDOT TECHNICAL ADVISORY COMMITTEE MEMBERS

The following WSDOT employees are members of the Technical Advisory Committee (TAC) for the project (as of June 09, 2008).

Main Contacts:

Kathy Lindquist
Research Manager
Office of Research and Library Services
(Primary contact)
lindquk@wsdot.wa.gov

Jim Culp
Interactive Communications Specialist
Communications Office
(Technical monitor)
culpj@wsdot.wa.gov

Other TAC Members:

Rebecca Christie, Librarian
Materials Laboratory
chrisre@wsdot.wa.gov

Cathy Downs, Manager
Records and Information Services
downsc@wsdot.wa.gov

Andy Everett, Data Catalog Administrator
Information and Metadata Architect
Office of Information Technology
everetta@wsdot.wa.gov

Mark Finch, Roadway Branch Manager
Transportation Data Office
finchm@wsdot.wa.gov

John H. Johnson, Records Management Coordinator
Records and Information Services
johnsjh@wadot.wa.gov

Leni Oman, Director
Office of Research and Library Services
omanl@wsdot.wa.gov

Kathy Szolomayer, Librarian
Office of Research and Library Services
szolomk@wsdot.wa.gov

Jim Walker, Manager
Aerial Photography Programs
walkerj@wsdot.wa.gov

Project Technical Advisor:

Richard Norrell, Support Supervisor
Enterprise Content Management (EMC)
Office of Information Technology
NorrelR@wsdot.wa.gov