Management of Spectral Imaging Archives for Scientific Preservation Studies

Doug Emery, Emery IT; Baltimore, Maryland/USA; Fenella G. France, Library of Congress; Washington D.C./USA; Michael B. Toth; R. B. Toth Associates; Oakton, Virginia/USA

Abstract

The Library of Congress is conducting preservation studies of historic documents, manuscripts and other cultural objects using advanced capabilities for spectral imaging, which poses challenges for the collection, storage and retrieval of data and associated metadata to accepted standards. Imaging studies are being conducted of the Waldseemüller 1507 world map, and other maps from the Library of Congress Geography and Maps Division, including L'Enfant's 1791 Plan of Washington D.C., the Carta Marina 1516 world map, and a range of daguerreotypes, as well as drafts of the Gettysburg address probably in Lincoln's handwriting. This imaging research is built upon spectral imaging of the Archimedes Palimpsest project. Based on these studies, the Library of Congress is developing a proposed standard for a customization of the Resource Description Framework for the semantic description and interchange of preservation reference material. This will define a core set of data elements required to identify and minimally describe a sample of reference material. The core data elements will need to provide methods for identification of and linkage to supporting files of multiple types: spectra, images, and documents. A proof of concept XML schema is being developed to help define best practices for adding further modules to the standard. With this metadata and data structure, libraries, archives and museums will not only be able to integrate spectral imaging as a useful tool for scientific and preservation studies, but be able to share data for effective storage, retrieval and collaboration to shared standards, nationally and internationally.

Library of Congress Preservation Research

The mission of the Preservation Directorate at the Library of Congress (LC) is "to assure long-term, uninterrupted access to the intellectual content of the Library's collections, either in original or reformatted form."[1] This can be challenging given that LC collections number over 121 million items, comprising iconography of more than two hundred years of US history, and including prior world historical artifacts.

Preservation of cultural heritage can be challenging, since all artifacts have deteriorated during their lifetime from usage and the environment. These deterioration processes can include the effects of light (both ultraviolet and visible), relative humidity fluctuations and moisture, temperature, oxidation, biological activity, pollutants and soiling, and historic treatments. Identification of specific compounds can be complicated due to the formation of deterioration breakdown products from exposure to a range of degradation effects over an artifact's history. This is compounded when much historical documentation about the artifact is frequently lost or not recorded. The interaction of compounds and products can change the response of the specific compound due to its interaction with other materials. This requires a select range of scientific analyses to understand the base material and accurately characterize it.

As part of its ongoing focus on developing non-destructive analytical techniques, the Preservation Research and Testing Division (PRTD) of the LC is conducting hyperspectral digital imaging research for a range of historic cultural manuscripts and documents and objects in its newly established optical imaging laboratory. These include the Waldseemüller 1507 world map and other maps from the LC Geography and Maps Division, including L'Enfant's 1791 Plan of Washington D.C. and the Carta Marina 1516 world map, the Nicolay and Hay handwritten drafts of the Gettysburg address (Fig. 1), an illustrated Armenian Gospel, Korean historic manuscript, Peruvian Harkness collection documents, textiles and a range of early daguerreotypes.

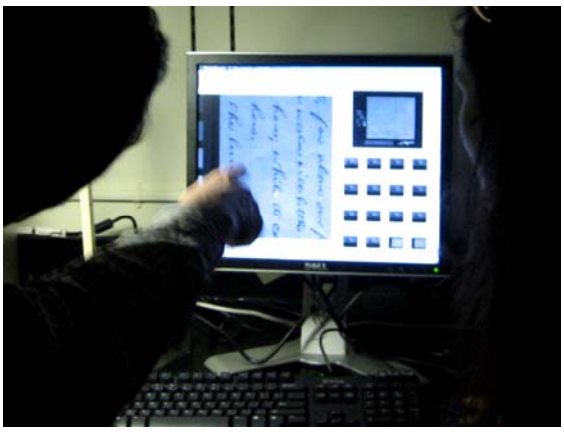


***Figure 1**. Real-time Text Analysis of the Gettysburg Address Nicolay draft*

Building on multispectral imaging of just a few broad wavelengths of light and previous work with ink analysis,[2] the LC's current hyperspectral imaging system measures a series of narrow bands of the visible and non-visible spectrum. Images collected may be digitally combined with or subtracted from each other to form processed images. These are used for precise analyses of a wide range of materials with specific spectral responses. This has the advantage of characterizing, identifying and possibly quantifying materials, and discriminating between similar compounds present in the document. The imaging is used to:

- "Fingerprint" documents with images from regions in both the visible and non-visible range,
- Create post-capture, high-resolution full color images, and

• Characterize components of the document.

Artifacts are made up of a range of material substrates, including paper, parchment, textiles, and photographic materials, upon which a range of media have been used to record text and illustrations. As part of characterizing components of these artifacts, specific spectral responses of inks, colorants, treatments and deterioration products are being analyzed, utilizing a low heat, low light, light emitting diode (LED) and integrated digital camera imaging system. To effectively characterize spectral components in a totally non-destructive manner, results must be compared with image files of known reference samples and files from comparative techniques to develop a library of spectral imaging reference samples.

Imaging a range of substrates and media provides a good example of the range of imaging required for different LC collection artifacts. The spectral imaging requirements are related to the specific artifact and its research questions. Building on prior research, the LC has found imaging paper documents often provides important characterizations of the inks and colorants, which can impact preservation of a document and treatment and display decisions.[4] As noted in the example in Table 1, if imaging helps identify areas where iron gall ink is present, due to the corrosive nature of this ink it is important to reduce oxidative and hydrolytic reactions. The imaging study of the Waldseemüller map not only characterized the ink and an iron gall inscription, but also identified certain colorants that are very susceptible to light and will benefit from the current anoxic environment. Other parchment or paper documents may have lost or hidden text and features that can be revealed through non-visible light. Imaging of the Nicolay and Hay drafts of the Gettysburg address with Lincoln's original text revealed text and corrections, while details of the L'Enfant Plan of Washington D.C. that were not visible to the naked eye were suddenly revealed.

Raking light images are also collected to reveal three-dimensional detail and topography. This was particularly useful with the Waldseemüller map, when the technique provided a view of how the original woodblock print may have appeared. Changes in baseline images can be compared digitally for evaluations of exhibition parameters, to observe changes in fragile items of significant cultural heritage. This was employed for gilded daguerreotypes, and will be for early un-gilded daguerreotypes that are very sensitive to light, requiring close monitoring of their time on exhibit. Imaging is currently contributing to LC preservation research of daguerreotypes in the Lincoln Bicentennial exhibit. "Before-and-after" spectral images provide information about the impact of exhibition conditions, a critical factor in the ongoing tension between preservation and access to cultural heritage.

Imaging paper and ink as part of baseline identification and calibration studies is critical for assessing the media and substrate and for treatment decisions. "Finger-printing" of documents with ultraviolet (UV) and infrared (IR) spectra allows the capture of information that is not apparent in the visible range, which can be useful for assuring the provenance of artifacts, as well as recovering information that is useful for researchers and scholars.

The optics laboratory has a range of equipment with many instruments that use proprietary software, requiring adaptation to standardized file readable formats. Analyses may include X-ray florescence (XRF), three-dimensional (3D) florescence, Fourier-transform IR (FTIR), environmental scanning electron microscopy (ESEM), and other chemical sensors. Integrating these file structures into a fully accessible preservation reference collection is the key to creating an internationally accessible data structure with integrity of data and a robust structured framework.

Spectral Imaging Program

Central to the collection of preservation information about the national and international items of significant cultural heritage is the hyperspectral digital imaging of collection documents and items. The LC uses a MegaVision-Equipoise imaging system for the efficient collection, illumination and storage of needed image data. This system was proven in the LC during the November 2007, Waldseemüller hyperspectral imaging project.[3] Illumination with EurekaLight™ LEDs yields precise spectral data for research and processing with low heat and light output. The light levels from a normal photographic imaging session are 3,000 lux or higher, while the light levels from this system have been measured at 3-5 lux. This is extremely important for reducing damage to fragile and degraded documents. This provides digital images of about 40 MB, which offers good resolution and file size for efficient processing and analysis on most computer systems. The MegaVision imaging capability offers an appropriate balance between resolution and processing and storage tractability with the following components (Fig. 2):

Imaging System: A MegaVision Monochrome E6, 39-megapixel monochrome back and camera, and PhotoShoot™ image capture software integrated with the illumination system. This includes a 2X3 View camera with Schneider Apo Macro 120 mm f5.6 lens.

Illumination: A pair of integrated EurekaLight™ 8-4 illuminators, each containing four UV-Visible and two IR subpanels with multiple LEDs in the following 13 wavelengths:

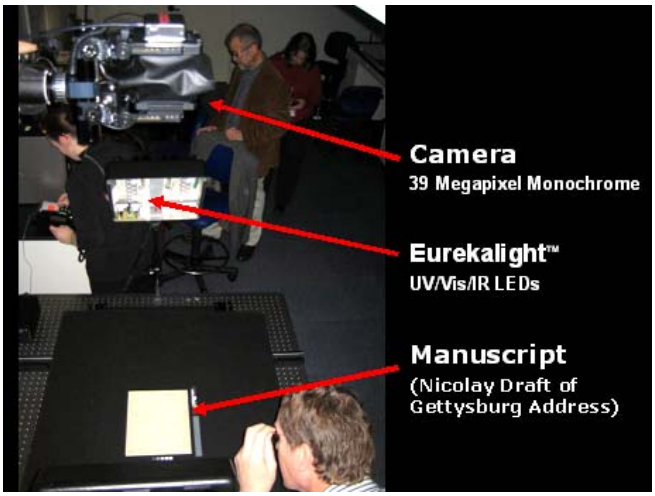
UV: 365nm; Visible: 445, 470, 505, 530, 570, 617, 625nm; and IR: 735, 780, 870, 940, 1050 nm.

Software: The imaging system is operated with MegaVision's PhotoShoot™ digital image capture software, a comprehensive image capture program that drives the LED illumination and the Schneider digital shutter and aperture in an integrated fashion. The software allows illumination to be customized, with each LED wavelength optimized for balanced distribution by setting individual LED exposure durations.

***Figure 2**. MegaVision 39 Megapixel Monochrome Camera, EurekaLights and*

Table 1: Impact of Identification on Preservation Decisions

Sample	Preservation Issue(s)	Identification Requirement	Preservation Treatment	Storage and Exhibition Impact
Iron gall ink ferrous sulfate + gallotannic acid Used to write on paper or parchment and vellum	Exposure to light, oxidation, and hydrolysis causes degradation due to high acidity and iron corrosion of collagen in and cellulose	Iron in ink suggests iron gall but not conclusive; therefore confirmation and identification of other compounds is necessary to complete the analysis.	Identification allows treatments to reduce oxidation and hydrolytic reactions by making the environment less acidic and retarding iron corrosion.	1. House in anoxic environment where the relative humidity can also be reduced and controlled. 2. Cold Storage 3. Deacidification



Nicolay Draft of the Gettysburg Address, Sheet 2

Images: In the LC application, this system generates an image cube of 16-20 consecutive registered images in ".dng" file format that is then standardized to a ".tiff" common file format for broader dissemination and study. One image cube from the LC imaging system yields approximately 1200 MB of TIFF files prior to any image processing or spectral combinations. Additional processing is performed for the creation of pseudo-color and principle component analysis (PCA) images to provide a common digital product for analysis, research, and exhibition, which is important for all artifacts that have been and will be imaged.

Data Management

To ensure the short- and long-term viability of hyperspectral imaging data, the LC employs and is refining processes and standards for data and metadata collection based on broadly accepted open international standards and models. Dublin Core is used for object identification and geospatial standards for spatially relating the data to the imaged object.[5] The LC is able to crosswalk the Dublin Core data to meet MARC standards.[6] For data formats, de facto standards (e.g., TIFF and XMP) or established standards (e.g., XML and EXIF) are used.

Data formats and standards are selected to ensure open access to and long-term viability of the data. Consistency in collection of data and metadata makes these goals achievable, helps to guarantee the intelligibility of the archive, and reduces the possibility of human errors that can corrupt data or impede access to it. As part of image planning, the LC develops goals and collection protocols that comply with these standards and consistent practice. Discipline is required during imaging to ensure that files are named consistently and logged, so that replicated and disseminated images all refer to the same source image and techniques. After imaging, images and log files are compared to correct inconsistencies in data collection and file naming, and complete metadata records are paired with the images. In future, imaging and metadata collection will be further integrated so that all metadata will be available during imaging and stored with the image files. The same requirements for data management and metadata collection apply to processed images.

The LC found that correct data management and data collection are critical processes for effectively collecting an integrated set of images for research – the core data product. Failure to address data management in a timely manner risks the loss of data, along with the likelihood that the imaging data will become unusable and/or unmanageable. The next step in data management is to move the data to a managed repository, such as a PREMIS system, which will ensure the integrity of the data and its accessibility over time.[7] Reference samples require metadata extensions that are compatible with the file format and structure and allow related data and research to be linked through the digital object.

Preservation Reference Information

The collection of hyperspectral imaging data is one portion of the larger effort at the LC to develop a body of information on cultural objects and preservation reference materials for accessible analyses to support preservation activities. This body of information adds to the LC's knowledge of the materials and components of cultural objects. The ultimate goal is to build databases of sharable data on objects and reference materials. All types of data – not just hyperspectral – available on a single object then will be accessible through a single source. The LC is bringing together data on cultural objects and material samples, including images, spectra, and testing results of all types, in a single capability for searching and sharing.

Preservation of collection artifacts at LC and other cultural institutions utilizes wherever possible non-destructive and non-invasive methods of analysis, including hyperspectral imaging. Creating a system where this can be undertaken with confidence and accuracy requires a range of analyses and comparison with accurately characterized base materials – including substrates, inks, colorants and treatments. As part of this effort, PRTD is establishing a collection of accelerated aged and naturally aged reference samples for comparison purposes. The naturally aged samples are invaluable in characterizing the deterioration that various artifacts have undergone, since these samples provide known information on substrates, inks, colorants and treatments that can then be compared with the data collected from original artifacts. The collection of this historical information detailing the exposure conditions of these samples is the critical component in this reference sample set. Capture of the metadata associated with each reference sample provides ready access to the necessary information. Non-destructive comparisons can then be made to the original cultural artifact, from which samples cannot be taken. This ultimately informs decisions about the longevity and preservation of the LC's collections of cultural heritage items. Sharing this information – nationally and internationally – in a usable format, while assuring the high quality and accuracy of the data, is critical for preservation professionals, especially when many institutions do not have access to a full range of scientific equipment, but can benefit from the analyses made elsewhere.

For accurate characterization of various components of an artifact, a range of non-destructive tests can be utilized and the collected data compiled and possibly directly linked to the document or object, through a "scriptospatial" mapping of the document. This effectively creates a spatial reference system for the data linked directly to points on the original object through the "digital object" itself. Due to deterioration of components through aging and exposure to a range of conditions, determining what constitutes characterization of the artifact by identification of

various compounds and components of the substrate and media can prove challenging. Therefore the reference collection needs to comprise data, images, and spectral files from the range of scientific equipment noted previously that may include spectral imaging, XRF, 3D florescence, FTIR, and ESEM, as well as other chemical and mechanical tests. These instruments characterize the organic and inorganic compounds that comprise the object by analyzing key factors, such as the spectral response and elemental analysis, and identifying the molecular structure and chemical groups present in the substrate and media. Part of the protocol for the reference collection data will be to ensure standardized file formats that are accessible with general desktop applications and are not restricted by proprietary file formats.

Metadata and Standards

For its hyperspectral imaging metadata, the LC employs the Archimedes Palimpsest Metadata Standard (APMS), which applies a geospatial metadata model to cultural objects.[8] Developed for the Archimedes Palimpsest imaging project, the APMS deals specifically with the data requirements to relate the multiple registered images in a spectral imaging "cube" to each other, to the scientific parameters of their creation, and spatially to the subject imaged: the map sheet, the manuscript folio, the textile, or the daguerreotype. The APMS defines six categories of information:

1. Identification Information
2. Spatial Data Reference Information
3. Imaging and Spectral Data Reference Information
4. Data Type Informatio
5. Data Content Information
6. Metadata Reference Information

Key to the effectiveness of hyperspectral imaging is maintaining the image "cube" of multiple spatially coterminous images of the same object. It is crucial for ongoing work in the LC to analyze the data by looking at the same region of the object across spectra. The description of the cube and the character and nature of each of the constituents are provided by the metadata. Respectively, the six types of information:

1. Uniquely identify the digital object (i.e. the image file);
2. Define the spatial relationship of the digital image to the imaged object using a Cartesian coordinate system;
3. Describe the imaging and illumination of the imaged object in objective scientific terms;
4. Detail the file type and the processes used to create it;
5. Relate the image to its physical source; and
6. Identify the metadata standards used to create the record.

The metadata are stored in the header of each TIFF image, using the TIFF "ImageDesciption" tag and using Adobe's XMP (Extensile Metadata Platform) format.[9] As XML frameworks for expressing structured data, RDF and XMP importantly allow for the expansion of the data model through the use of XML namespaces.[10] In the case of XMP, the implementation specifies default namespaces that apply to the file type. These include, for example, Dublin Core, rights management, and, for images, TIFF and EXIF. The system is designed to accept additional data types with the addition of other XML namespaces. This makes XMP ideal for imaging data standards like the APMS, which includes Dublin Core elements and overlaps with TIFF and EXIF fields. In the case of the APMS, the data model includes a custom namespace and schema for information specific to the standard.

Development of RDF Standard

To extend the knowledge of the condition and content of documents and other objects within and outside the LC, PRTD is working to preserve and access the digital data with its metadata, and maintain links between the new data and associated information reference data for the object. It is developing an open framework for the images and reference collection that can be made internationally accessible with a standardized file format. The Resource Description Framework (RDF) has been identified as a method for bringing together associated structured data of different types.[11] It will allow establishment of a model that has a core set of data types, including images, to allow others to add other structured data for specialized needs. In the LC this will allow core object identification and characterization to be added or linked to hyperspectral imaging data or testing spectra.

With XMP as a model, the LC is working on this format as an implementation of RDF, starting with spectral images of the L'Enfant plan of Washington D.C. The key problems with the data model are 1) identifying and characterizing objects and samples so that they may be discovered and referenced, and 2) accommodating other types of metadata required to document and locate the products of testing and research. The first component, "identification and characterization," forms the core namespace of each record and provides a common point of access for all records. This component will provide the location of the object, intellectual property and institutional responsibility information, the type of the sample, its substrate and other contextual information, a categorization of the object or sample using controlled vocabulary, and a unique identifier. As much as possible, existing systems and standards for identification and characterization are used. For example, the Library of Congress Control Number (LCCN) for a daguerreotype, scientific description of its components, Dublin Core metadata and comprehensive keyword assignment from a controlled vocabulary form the core of the identification and characterization of this object and its associated spectral images.

The inclusion of information about test and research data with the image data poses a special problem. While RDF does allow for the expansion of records through other types of data, as with XMP, it is not a metadata wrapper like METS, which can store any type of namespaced XML data as well as embed text-encoded binary data, like image files.[12] The goal, however, is not to encapsulate all scientific data about a sample in a single file. The size of the data may not be known prior to access, and may be prohibitively large. Some imaging data sets on a single object are multiple gigabytes in size. Rather the goal will be to permit discovery of the object – e.g. all pigments of a certain color, or material type, or substrate – and describe and provide links to all available test and research data – imaging studies, or XRF or FTIR spectra. The research data component of the data model will provide summary information of external data, technical information adequate for understanding the type of research data, and pointers to the file or files that belong to the test data. The RDF core identification of a sample with hyperspectral imaging data, for instance, contains a set of elements in their own namespace to reference hyperspectral data, and the component parts of such a set, so that users could discover and request data of interest to them. A task of the development of the RDF metadata data model will be to determine the amount of specialization required to characterize and provide links to different types of testing and research data. The resulting

system should provide services for browsing and querying based on defined criteria, return standard metadata records, and allow the retrieval of requested scientific results. Coordination of research data is enhanced through the ability to search across the datasets.

Conclusion

The body of preservation knowledge required to protect and preserve our history will be expanded with access to a data set of spectral images and metadata from a range of reference samples that have been created or collected in conditions that represent or replicate those to which original artifacts have been exposed. This allows analysis of risks to an artifact's preservation based on scientific studies of the mechanisms involved. The broad base of scientific data then contributes to implementation of objective preservation decisions. The development of a structured, open access architecture for the sharing and advancement of preservation research data is critical to preservation of a range of items of significant cultural heritage in the LC and around the world. In the LC this is based upon accepted imaging and data standards – a critical component in the reference collection that serves as a baseline for non-destructive preservation analyses. Only with the implementation of standardized file formats, structures and well-developed metadata protocols can this open access system remain robust and retain the data integrity critical for cultural heritage preservation. The utilization of RDF to collect and bring together different types of structured data allows the expansion of core data about the images to accommodate a range of preservation data. Scriptospatial mapping can then link these various data to the original object, as well as the digital image product. The reference materials metadata standard is a customization of RDF that can then define the core data as well as expand it to allow for both the semantic description and interchange of preservation reference data and materials. Increasing our understanding of deterioration mechanisms and how these affect various substrates such as paper, parchment, textiles, and fragile photographic media enhances the development of non-destructive testing. This is greatly enhanced with access to characterized reference samples. The development and characterization of reference materials to allow broad access is a powerful capability to enable the effective global exchange of scientific data for collaboration and development of preservation knowledge.

Acknowledgments

The authors want to thank the team of preservation and imaging scientists and conservators, whose support for imaging and the artifacts make this continuing research possible at the LC and other institutions dedicated to cultural heritage preservation.

References

[1] Library of Congress, "Mission of the Preservation Directorate" 16 Oct. 2006, <http://www.loc.gov/preserv/mission.html> (6 Dec. 2008)

[2] Knox, Keith T., et. al Image Restoration of Damaged or Erased Manuscripts, European Signal Processing Conference, Lausanne http://www.eurasip.org/Proceedings/Eusipco/Eusipco2008/papers/1569105284.pdf (2008)

[3] Christens-Barry, W.A., et al, Camera system for multispectral imaging of documents, Proc. SPIE, pg. 724908-1 - 724908-10 (2009).

[4] Suarez, A.V., Tsitsui, N.D., "The value of museum collections for research and society" BioScience, vol 54, No. 1, January (2004)

[5] Dublin Core Metadata Initiative (DCMI), Dublin Core Metadata Element Set, Version 1.1., <http://www.dublincore.org/documents/dces/> (2008)

[6] Library of Congress' MARC Standards, 13 September 2008, <http://www.loc.gov/marc/> (2 Dec. 2008)

[7] Guenther, Rebecca S.; Angela Dappert, Markus Enders, "Using METS, PREMIS and MODS for Archiving eJournals," D-Lib Magazine, 14:9/10, September/October (2008), <http://www.dlib.org/dlib/september08/dappert/09dappert.html>

[8] Archimedes Palimpsest Program, Archimedes Palimpsest Metadata Standard 1.0, Revision 5. Baltimore, Maryland: Walters Art Museum, 7 June (2006)

[9] Adobe Systems Incorporated, Extensible Metadata Platform (XMP) Specification: Part 1, Data and Serialization Model, 2008

[10] Extensible Markup Language (XML) 1.0 (Fifth Edition) W3C Recommendation, 26 November 2008 <http://www.w3.org/TR/2008/REC-xml-20081126/> (2 Dec. 2008)

[11] RDF/XML Syntax Specification (Revised) W3C Recommendation, 10 February 2004 <http://www.w3.org/TR/rdf-syntax-grammar/>

[12] Library of Congress Network Development and MARC Standards Office, Metadata Encoding and Transmission Standard, 13 September 2008, <http://www.loc.gov/standards/mets/> (1 Dec 2008)

Author Biography

Doug Emery *is the Emery IT Data Manager responsible for the image metadata collection, data storage and distribution of the Archimedes Palimpsest and other projects. His educational background is in ancient cultures and languages, but he now works as a programmer and database administrator in academics and private industry. Doug applies professional expertise in data management and international metadata standards, as well as an understanding of the issues involved, in working with ancient manuscripts.*

Dr. Fenella G. France *is a Preservation Scientist at the Library of Congress. An international specialist on environmental deterioration, she researches non-destructive imaging techniques, as well as anoxic and protective environments, including cases for significant cultural heritage items. She received her PhD from Otago University, New Zealand, and MBA from Deakin University, Australia. As the research scientist for the Star-Spangled Banner project for the Smithsonian Institution, she determined the exhibition requirements for this US icon.*

Michael B. Toth *is President of R.B.Toth Associates, providing management, systems integration and strategic planning for the integration of technical systems to the study, preservation and display of cultural objects in museums and libraries. Mr. Toth brings extensive experience with his work on advanced information and space systems for the US Government. For over 20 years he has managed the development, integration and operation of global imagery, IT and GIS systems.*