https://doi.org/10.2352/issn.2168-3204.2018.1.0.14

Provenance-Oriented Documentation of Multi-spectral Data

Ya-Ning Chen; Department of Information and Library Science, Tamkang University; New Taipei City, Taiwan

Simon C. Lin; Institute of Physics, Academia Sinica; Taipei, Taiwan

M. James Shyu; Department of Information Communications, Chinese Culture University; Taipei, Taiwan

Eric Yen; Centre for Information Technology Innovation, Academia Sinica; Taipei, Taiwan

Background

Nowadays scientific data can be generated and collected very easily with advanced instruments and facilities. In recent years, scientific data or datasets have emerged as an additional important source of scholarly output. Data-centric research needs data as a catalyst to inspire new research by repurposing or combining existing research data. Currently, most primary scientific data lack good documentation and management for future reuse, and are usually locked in personal archives as a part of dark data/archives that are in danger of being lost. Although data documentation is tedious and labor-intensive, data documentation is still cheaper than data reproduction or recollection [1].

Problem

In pursuit of knowledge, scientists design experiments or make observations using scientific instruments and related techniques to investigate specific research questions. During the research lifecycle, various research data are generated with notes written in laboratory notebooks at different stages. This content always includes a lot of personalized codification and scratched graphs and notes, and cannot be understood and interpreted correctly by outsiders or a third party. Although various approaches, such as data papers and metadata, have been proposed for the documentation of research data, they all focus on finalized research data from a specific stage in the whole research lifecycle. In fact, at different research stages different kinds of research data are generated for distinctive purposes such as data creation, comparison, analysis and publication. Not all the various research data at different stages of the research lifecycle are collected for documentation. Furthermore, the relationships (e.g., data derivation, version and causal-effect) between data at different research stages that illustrate the data lineage are not described to indicate the provenance of data. Furthermore, some data are difficult to reproduce and recollect due to the high cost of the re-acquisition of research data. More importantly, some data cannot be reproduced or recollected, such as observational data of extinct species. Appropriate documentation is a cornerstone for the sharing and reuse of research data. Therefore, investigation of best practices for data description that can be applied to research data in a practical way to facilitate sharing and reuse within a scientific community is valuable.

Approach

In this study the "Taiwan-Mongolia Joint Project: The Innovative Application of Multi-Spectral Technology for Culture Heritage" (TW-MN Joint Project) was selected as a case study to illustrate methods to document various types of multi-spectral data to facilitate discovery, sharing and reuse, through a provenance-oriented metadata approach. The TW-MN Joint Project is a collaborative project between the Mongolian Academy of Sciences, and Academia Sinica, the Chinese Cultural University and Tamkang University in Taiwan. This project aims to assist Mongolian researchers by building their first pilot case adopting multi-spectral imaging technology for cultural heritage. The goal of this project is to initiate a new era for the digital humanities community and digital restoration community by taking advantaging of a mature multi-spectral technology and analysis model.

First, a path-based analysis approach was adopted to analyze the relationships between various multi-spectral data to illustrate the data lineage in terms of provenance. This means that various research data generated by TW-MN Joint Project were collected at different stages of the research lifecycle in a sequential way. Each type of research data is regarded as an independent node and then their relationships are depicted and connected by directional arrows to build up a path-based graph for illustrating the provenance between the research data. Second, an interview was conducted with a principal investigator (PI) to collect and analyze the aforementioned project requirements to develop metadata elements to describe the contextual information of multi-spectral data belonging to each stage and the provenance relationships between data stemming from various stages of the research lifecycle, such as the environmental settings for generating research data and related information about processing the multi-spectral data. The Core Scientific Metadata Model (CSMD) [2] developed by the Science and Technology Facility Council (STFC) in the UK was used in this study. The reason for integrated

adoption of the CSMD was two-fold: first, the CSMD is to develop metadata elements for research projects and its derived data in terms of research lifecycle, and second, scholarly publications including presentation slides and journal articles were also included in this study as subject in addition to research datasets. Thus, it is reasonable to not only directly borrow metadata elements from CSMD, but also allow the project manager to develop new metadata elements for local requirements of The TW-MN Joint Project according to the principles of application profile [3]. Finally, the Comprehensive Knowledge Archive Network (CKAN) platform, which is open source software for managing research data, was employed as a test-bed to examine the feasibility of the proposed provenance-oriented metadata approach.

Results

In terms of stages within the research lifecycle, five types of research data were generated for the TW-MN Joint Project as follows: uncalibrated, for primary data; calibrated, for processing data; LAB, for visualization, application for printing output; and analysis data, for scholarly publications such as journal articles. Applications for printing output can be separated into various subtypes for the following purposes: sRGB (standard red green blue) for computer screen, Adobe RGB for general printing, and CMYK (cyan, magenta, yellow, and black) for professional printing. Analysis of multi-spectra data is often used for figures, graphs or supportive evidence in journal articles. In terms of path-based provenance, two kinds of relationships exist between the aforementioned five types of research data. One is the sequential path-based provenance relationship between un-calibrated, calibrated, LAB and analysis data in an ordered way. The uncalibrated data is the primary research data for this project. This means that the uncalibrated data is the original data for other types of derived multi-spectral data with different processing and purposes. The other relationship is the splitting path-based relationship between calibrated and its derived data, including LAB, versions of RGB computer screen and Adobe RGB, and analysis data. In other words, calibrated data can be processed to fit various purposes, including visualization, human recognizable versions and scientific analysis (Fig. 1). Generally, formal scholarly publications are excluded from research on data management. However, in this study, formal and informal scholarly publications and their derived works such as presentation materials have been included as a sixth type of data output to illustrate the complete activities of the project-based research lifecycle. In terms of functional application, the aforementioned six types of data can be re-categorized into six types of data as follows: device-dependent raw image data, device-independent calibrated spectral image data, visualized image data, application image data, analysis data and scholarly publications (Fig. 2).

In terms of documentation, two objectives for functional requirements for a provenance-oriented metadata approach to data documentation were achieved as follows: contextual information and provenance. First, in this study metadata elements were borrowed from the Core Scientific Metadata Model (CSMD) [2] to describe the attributes of the aforementioned five types of research data and scholarly publications in the TW-MN Joint Project. Next, all borrowed metadata elements were filled with instances for each type of research data and scholarly publications. In terms of functional application, the focal points of the description of metadata for the four types of multi-spectral image data are as follows:

- Device-dependent raw image data: including hardware configuration and how the signal is generated, reference material, what kind of the reference target and how the reference data, sequence and orientation of the bit stream, and colorimetric measurement data.
- Device-independent calibrated spectral image data: including the range of the spectral signal and bandwidth for each sampling.
- Visualized colorimetric image data: including the spectral power distribution of the light source and the viewing distance to the object.
- Application image data: process method, signal encoding format, file format and color mode.

In addition to the significant attributes of multi-spectral image data, information related to project, investigation, datasets, data files and scholarly publications were also included as metadata elements for documentation. Under the principles of application profile [3], most metadata elements were selected from CSMD directly. Only two new metadata elements were created and one was redefined for local requirements of the TW-MN Joint Project as follows: processing information and reference information (Fig. 3). Based on confirmation by the PI, two tiers of key metadata elements were generalized to restore essential contextual information related to original environmental settings for the generation of research data. The metadata elements of the first tier are nearly shared by all types of research data, and those of the second tier are applied to specific types of research data (Fig. 4 and Table 1). Third, four-level metadata elements which are similar to fonds, series, files and items in archives were implemented into the CKAN platform to indicate the provenance relationships between types of research data from top to bottom as follows: project, investigation, dataset and data file (i.e., organization, group, dataset and resource in CKAN respectively). Therefore, provenance and contextual information related to multi-spectral image data were documented to emulate for restoring all the environmental settings for generation of research data to facilitate data sharing and reuse in the future.

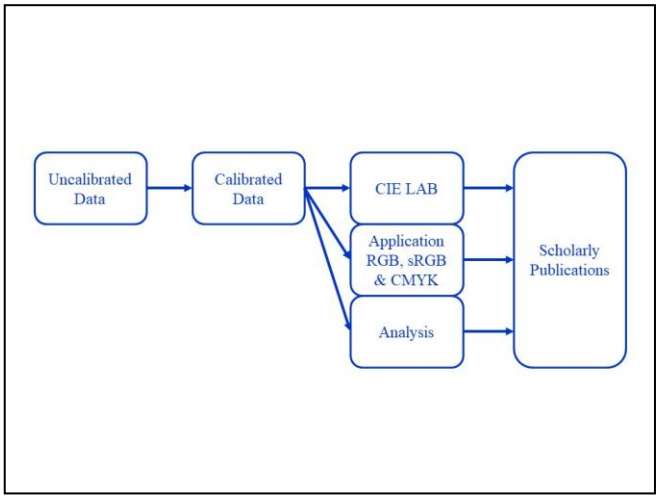

Fig. 1 Provenance of multi-spectral image data of the TW-MN Joint Project in terms of data type.

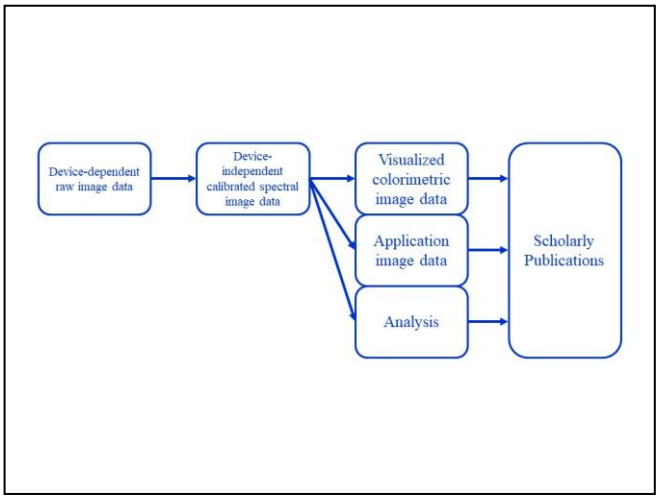

Fig. 2 Provenance of multi-spectral image data of the TW-MN Joint Project in terms of functional application.

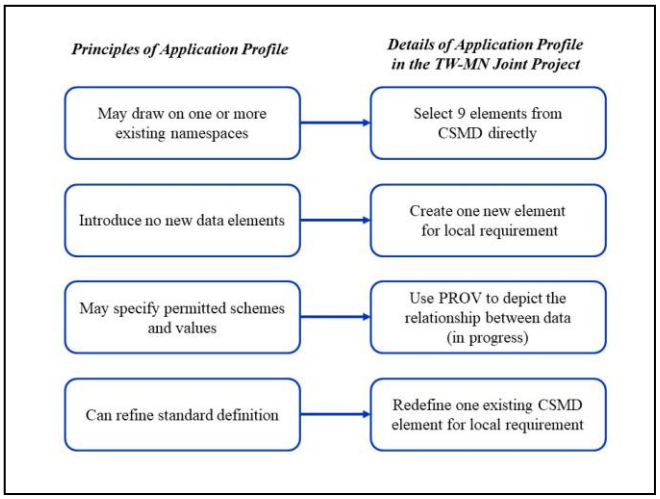

Fig. 3 Implementation of application profile into metadata design for the multi-spectral image data of the TW-MN Joint Project.

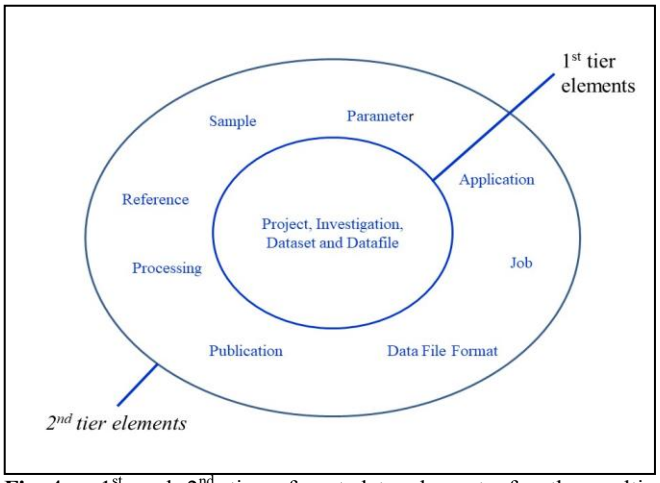

Fig. 4 1st and 2nd tier of metadata elements for the multi-spectral image data of the TW-MN Joint Project.

Table 1 Metadata elements for multi-spectral image data of the TW-MN Joint Project.

Element	Sub-element	Source	Tier
Project	Name, period, objective, and identifier	CSMD	1st
Investigation	Title, abstract, dates, and identifier	CSMD	1st
Agent	Funding organization	CSMD & redefinition	1st
Facility	Name and description	CSMD	2nd
Instrument	Name and description	CSMD	2nd
Sample	Name, type, and type name	CSMD	2nd
Parameter	Type name, type description, and type unit	CSMD	2nd
Dataset	Name, type, version	CSMD	2nd
Datafile	Name, type, version, format	CSMD	2nd
Reference	Material, target, and data	New	2nd
Processing	Method, color mode, signal encoding, and sequence and orientation of the bit stream	New	2nd
Publication	DOI and full reference	CSMD	2nd

Conclusions

According to the provenance-oriented metadata approach, two tiers of metadata elements are not only used to describe the original environmental setting and related attributes of generation of research data, but provenance of research data is also kept to illustrate the data lineage based on the results of path-based analysis. Although the proposed approach can capture key information for each type of research data at various stages of the research lifecycle, more case studies are still needed to enrich the scope and types of the proposed approach. For example, distinctive path-based provenance and multi-spectral data may be generated for different purposes of scientific investigation. Furthermore, the PROV ontology and data model [4-6] provided by W3C has been adopted by CSMD, but no existing classes and properties have been implemented into CSDM, Therefore, it is worth examining the feasibility of illustration of provenance relationships between multi-spectral image research data and move forward to linked data.

References

[1] V. Chavan, and L. Penev, "The data paper: a mechanism to incentivize data publishing in biodiversity science," *BMC Inform.* **12**, S15: S2. https://doi.org/10.1186/1471-2105-12-S15-S2. (2011).

[2] B. Matthews, and S. Fisher, "CSMD: the Core Scientific Metadata Model," 2013, http://icatproject-contrib.github.io/CSMD/CSMD-4.0.pdf (25 October 2017)

[3] R. Heery, and M. Patel, "Application profiles: mixing and matching metadata schemas," *Ariadne* **25** (2000). http://www.ariadne.ac.uk/issue25/app-profiles (28 October 2017)

[4] Y. Gil, and S. Miles. Eds., "PROV model primer," 2013, https://www.w3.org/TR/2013/NOTE-prov-primer-20130430/ 3 March 2018)

[5] T. Lebo, S. Sahoo, and D. McGuinness, D., Eds., "PROV-O: the PROV ontology," 2013, https://www.w3.org/TR/prov-o/ (3 March 2018)

[6] L. Moreau, and P. Missier, Eds., "PROV-DM: the PROV data model," 2013, https://www.w3.org/TR/2013/REC-prov-dm-20130430/ (3 March 2018)

Author Biography

Dr. Ya-Ning Chen is an Associate Professor in the Department of Information and Library Science at Tamkang University in Taiwan. He pioneered metadata research and service, supporting metadata implementation and analysis of over 100 theme-based projects from the National Digital Archives Program (NDAP, 2002-2007) and the Taiwan e-Learning and Digital Archives Program (TELDAP, 2008-2012) in Taiwan. His research interests are information organization, digital libraries, open access, linked data, research data management and social networks.

Dr. Simon C. Lin is the Project Director of the ASGC (Academia Sinica Grid Computing Centre). Under his leadership, ASGC has assisted 30 sites in Asia to join the global e-Science collaboration from High Energy Physics, Biomedicine, Earth Sciences, Climate Change, Digital Archives, etc. Dr. Lin also pioneered the Digital Library/Museum Pilot Project which later led to the TELDAP. Within ten years, TELDAP has generated 8.8 million digitized materials ranging from anthropology to zoology.

Dr. Ming-Ching James Shyu is a professor in the Department of Information Communications, Chinese Culture University, Taipei, Taiwan. He received an M.S. in Color Science from Rochester Institute of Technology and Ph.D. in Color Imaging from Chiba University. Recently he developed a multi-spectral imaging system to capture heritage objects. He is Taiwan's national representative to the CIE Div. 8 and a member of several CIE technical committees. His research interests are in color imaging and photography.

Mr. Eric Yen is a Senior Research Scientist at the Grid and Scientific Computing Centre of Centre for Information Technology Innovation in Academia Sinica. He has more than 15 years' experience in distributed computing systems such as Grid and Clouds and digital archives. He works on multispectral data information system and analysis development in collaboration with researchers in Mongolia.