

# Empirical Humanities Metadata

## Working Group

### Case Statement Proposal

---

#### CASE STATEMENT PROPOSAL FOR THE EMPIRICAL HUMANITIES METADATA WORKING GROUP

#### Contents

1. WG Charter	1
2. Value Proposition	1
3. Engagement with Existing Work in the Area	3
4. Work Plan	4
5. Adoption plan	6
6. Initial Membership	7
7. References	9
8. Appendix A: Leadership Biographical Notes	9

#### **1. WG Charter**

The empirical humanities include history, folklore, cultural anthropology, and other fields in which researchers collect primary data of different types that can be used for cultural analysis. Today, these researchers often need to collaborate to understand phenomena that operate across geographic regions, scale, and communities of people. Established research practices and infrastructures in the empirical humanities do not support this collaboration, and our preliminary ethnographic research shows that anxiety over metadata practices is one of the key barriers limiting data sharing and cooperation.

The Empirical Humanities Metadata WG (EHM) will conduct research and develop an adoptable and adaptable protocol<sup>1</sup> for designing metadata management plans (MDMPs) in empirical humanities projects. A metadata management plan (MDMP), as a subcomponent of an overall data management plan (DMP), will be a formal document that outlines how metadata elements will be selected, structured, implemented, and managed within a project, and how social conventions will be developed to ensure discoverability and citability of data. The protocol will include a set of questions to ask while generating or evaluating MDMPs, and a collection of resources and guidelines for answering those questions in different ways depending on the needs of specific projects. This question and answer workflow, developed with input from an advisory board, will guide researchers through decisions about standards, policies, and cultural expectations to make ethnographic and historical data archivable, discoverable, and shareable.

#### **2. Value Proposition**

---

<sup>1</sup> We use the term protocol here to mean a recommended procedure. We also embrace the connotations of modifiability and iterability of protocol as “original draft” of a document, creating a foundation for further deliberation and adaptation.

### *Need for Collaborative Empirical Humanities in the Contemporary Moment*

Given the complexity (cultural, social, technical, ecological, economic, etc.) of many global problems today, collaborative empirical humanities research has renewed urgency. The interpretive, or hermeneutic, layer foundational to the empirical humanities is essential for navigating multi-faceted global problems. Drawing together diverse perspectives is also increasingly important. For decades, research in the empirical humanities has been an almost entirely individual-centric enterprise. Field notes, found documents, found or researcher-created photographs or recordings and other data used in cultural analysis are very rarely shared, except when reduced or rendered into some form of publication or museum display.

### *Navigating the Metadata Landscape*

One of the primary barriers to sharing data within the empirical humanities is a lack of agreed-upon protocols for choosing metadata standards for user-created primary research data. While there has been a great deal of work in the cultural heritage arena, especially within museums and libraries, and the dilemmas of qualitative data reuse are well documented (see Holstein and Gubrium 2004), the issues associated with preparing data for later use by third parties are yet to be thoroughly conceptualized. One challenge is the proliferation of metadata standards in recent years. In the cultural heritage sector, for example, Jenn Riley has identified 105 metadata standards and notes that “the sheer number of metadata standards in the cultural heritage sector is overwhelming, and their inter-relationships further complicate the situation” (Riley 2009). In contrast, the RDA Metadata Standards Directory WG lists only one standard for heritage studies and one for anthropology.<sup>2</sup> Many researchers find themselves caught in the confusing space between the dizzying proliferation of standards and a one-size-fits-all approach that can miss out on the diversity of data practices within disciplines.

### *Metadata Standards for Epistemological Pluralism*

Building digital infrastructure to support more data sharing and collaboration in the empirical humanities is far from straightforward. Analytic techniques in the empirical humanities differ from those in social science fields that may collect similar data, and are more akin to those used in literary and philosophical research, relying primarily on hermeneutics (interpretation for explanation and evocation rather than representative or statistical sampling for identification and validation). The goal is not to develop a concise and consistent view of an object, but to produce and explore multiple views of an object, leveraging “epistemological pluralism” (Keller 2002; Turkle and Papert 1990). Indeed, providing multiple, different interpretations and ways of representing particular phenomena (the socio-cultural causes and impacts of the Fukushima nuclear disaster, for example, or the impact of genetics research on understandings of environmental health) is the *key task* for empirical humanities researchers. Computational advances that support open-ended, underdetermined engagement with digital content that enables (even encourages) drift and transmutation in the way content is identified and taken up in analysis, are thus required.

Metadata is particularly complex and dynamic in the empirical humanities, even more so when research is collaborative. Empirical material often has limited or contested provenance information; the “empirical” itself can shift, in relevance or prevalence, as analytic structures

---

<sup>2</sup> <http://rd-alliance.github.io/metadata-directory/subjects/>, accessed May 5, 2017.

evolve and multiply. Qualitative interviews are not just collected, for example, they are *produced*, through questions and other elicitation techniques developed by the interviewer (often drawing on complex traditions of thought about language, culture, and society). Interviews are then analyzed, again using analytic structures developed within complex traditions of thought. If interviews are analyzed collaboratively, different analytic structures may be used by different researchers, or different researchers may deploy “the same” analytic structure in different ways, and come to different interpretations of what an interview, image or document “says.” It is thus critical to recognize – and make accessible and discoverable (if researchers deem this appropriate) – the analytic structures through which data in the empirical humanities is both produced and interpreted. Metadata functionality thus needs to be in place at many stages in the ethnographic research process, addressing diverse types of “data”—including analytic structures used to produce and interpret empirical data. This makes selecting metadata standards and structuring and managing metadata in an empirical humanities project particularly challenging. The proposed Working Group will help empirical humanities researchers overcome these hurdles by offering resources that suggest recommended procedures for implementing metadata in their projects.

Working closely with existing metadata-focused RDA groups (providing feedback on the list of elements in the package presented at the WG Chairs Meeting in Gaithersburg, for example), we will produce an empirical humanities-focused protocol for writing and implementing metadata management plans. The protocol will offer recommendations for content that should be included in a metadata management plan such as how to identify and adapt metadata standards, how to implement and manage metadata, and how to make data citable and discoverable. Once endorsed by the RDA, and taken up by early adopters, these recommended practices will be a go-to resource that researchers can adapt to meet the needs of their projects. Documenting metadata management plans will make user-created research data more findable and usable within these research traditions.

Individuals, communities and initiatives that will benefit from the proposed WG include:

- *Researchers*: by reducing psychological, institutional, political, cultural, and technological barriers to digitizing and sharing data, making shared data easier to find and cite, and improving mechanisms for credit
- *Collaborative research platform developers*: by providing a guide to various metadata options and recommendations about how metadata should be selected, structured, managed, and made discoverable
- *Interlocutors*: by better documenting the provenance of interview data, addressing questions of data ownership and confidentiality
- *Collections*: better metadata will improve accessibility, raising demand for archived material, helping collections better fulfil their mission.

### **3. Engagement with Existing Work in the Area**

Our preliminary research suggests that researchers in history and ethnography can quickly become overwhelmed by the multitude of diverse and somewhat scattered metadata standards and that many researchers have their own ideas about the limits of existing metadata practices and standards. Historian of cartography Pat Seed, for example, is involved in efforts to define

best metadata practices for maps and has noted that Dublin Core, recommended by many researchers she has sought advice from, is far from sufficient for her purposes. Another common standard, the Open Archives Initiative (OAI), informs the metadata fields used by many advanced digital projects supporting historical and ethnographic research, but one researcher we interviewed suggested that the OAI standards are "out-of-date" and that "the standard uses older web technology and has not been updated or changed in quite some time... if I were to do it today, first we would want a separation of the data model and the data encoding. For example, many APIs allow you to get results back in JSON, XML, RSS, etc." This WG will outline a protocol for questioning the relevance and appropriateness of existing standards and conventions and then will point to relevant literature for answering questions and adapting standards for particular purposes.

We will partner with existing RDA Groups, such as the Metadata IG, the Metadata Standards Directory WG, the Research Data Provenance IG and the Engagement IG. Individual researchers and groups within the RDA working on linked data, preservation, persistent identifiers, dynamic data citation and the long tail of research data will also be key partners. These connections have been strengthened at the RDA Ninth Plenary. Our engagement with RDA, primarily through the Digital Practices in History and Ethnography IG (DPHE) since 2013, has also brought us into contact with many organizations that support digital practices in the humanities and are already thinking about recommended practices for metadata, including DARIAH, ICSPR, CLARIN, Ariadne and Parthenos, and Europeana and Europeana Cloud. Beyond the RDA, we will engage with institutions with widely respected standards (i.e. the Smithsonian), initiatives (i.e. Open Folklore) and publishing bodies with a digital presence (i.e. the Journal of Cultural Anthropology). We will appoint members of these (and more) institutions to an advisory board, drawing on their expertise and perspectives in consultations at various stages of our work flow, including review of a beta version of the protocol.

#### **4. Work Plan**

##### *Work Plan Components*

##### **1. Preliminary Research (discussed at Plenary 9 and reported on during follow-up call-ins)**

- a. *Survey of relevant literature and projects* in order to build initial, and draw on existing, use-case scenarios for how metadata practices are developed and evaluated in the empirical humanities.
- b. *Ethnographic interviews* with researchers in history and ethnography, as well as other digital empirical humanists, on the types of data for which they need metadata practices, the scenarios in which they encounter metadata decisions and (with a focus on interviews and field notes) their practices.
- c. *Engagement with multiple existing RDA initiatives* including the Metadata IG, the Data Fabric IG and the Repository Platforms IG. We hosted an "issues share" call-in on metadata with representatives from the Metadata IG and other RDA groups and have received positive feedback from leadership within metadata-related RDA groups on this case statement and the complementarity of the EHM and these existing initiatives. We also held a joint session at Plenary 9 to

plan for collaboration between the Metadata Standards Directory Working Group and the initiative proposed here.

- d. *Preliminary work conducted within the DPHE suggests that* empirical humanities researchers often struggle to recognize and address metadata concerns such as:
    - i. When should a project use metadata standards specific to our domain, and when should a project use more general standards?
    - ii. Should metadata be nested or flat?
    - iii. Should a project be using controlled vocabularies, and if so, which?
    - iv. Should naming standards govern how metadata is implemented in collaborative projects, and if so, which?
    - v. How should a “data dictionary” be structured (with respect to the epistemological pluralism that characterizes much collaborative humanities work)?
    - vi. How should collaborative projects coordinate metadata implementation amongst dispersed researchers?
    - vii. Through what mechanisms will metadata be extractable from the project?
    - viii. How should a project use metadata to cite materials?
2. **Drafting Protocols and Initial Implementation (to be reported on at Plenary 10)**
    - a. *Drafting a beta protocol* for developing a new (or evaluating an existing) metadata management plan, with a focus on asking the right questions.
    - b. *Drafting and implementing a metadata management plan* for PECE (see WG Operation below) that responds to questions we’ve identified in surveys and interviews.
    - c. *Outreach and advisory board formation*, with at least ten members, leveraging contacts already made through DPHE-IG.
  3. **Review and Iteration (to be reported on at Plenary 11)**
    - a. *Reviewing the lessons learned* in implementing the protocol to create the PECE metadata management plan.
    - b. *Reviewing and updating metadata management plan protocol* with input from the advisory board and WG membership.
    - c. *Facilitating uptake* of protocol for developing or evaluating metadata management plans with adopters.
    - d. *Collecting and responding to feedback* on the protocol from early adopters.
  4. **Reporting on Conclusions and Sustainability Plan (Plenary 12)**
    - a. *Reporting on lessons learned* in initial uptake, and working with the DPHE-IG to ensure sustainability and evolution of the deliverables and their uptake.
    - b. *Promoting* the deliverables from this WG within the RDA and beyond.
    - c. *Coordinating* with the Metadata Standards Catalog WG to begin planning a new working group to make the protocol for developing MDMPs relevant to other domains and integrating the protocol in existing metadata and data management tools (such as the DMP Tool and the Metadata Standards Catalog)

#### *WG Operation*

The initial core members of this WG will meet regularly to ensure continual development towards the proposed deliverables. Many of the initial WG members have a well-established

working relationship, with a record of collaborative peer-reviewed publications and presentations (at the American Anthropological Association, the Society for the Social Studies of Science, the Society for Cultural Anthropology, and other conferences) as well as public dissemination (blog posts, university press releases, etc.) disseminating the results of their work. Vivian Wong, an RDA fellow working closely with the DPHE-IG and archivist with formal training in metadata, will support this work. Differences of opinion and experience will be viewed as an asset within this WG, and will be resolved through communication and collaboration practice.

In the spirit of “user-centered design,” this WG will partner with developers of PECE from the early stages to increase the likelihood of the deliverables meeting user needs. An ongoing series of “project shares” and “issues shares” hosted by the DPHE-IG (ongoing since 2013) will also provide frequent opportunities for members of this WG to envision how the deliverables could feed into a wide variety of digital humanities projects. This WG will be a vehicle for the broadly understood need for RDA to continue developing engagement with social science and humanities research communities.

In the second phase of the work, we will appoint an advisory board of at least 10 members with expertise in metadata standards and conventions, the empirical humanities, and archival procedures. We will hold an initial meeting with the full advisory board to go over the scope of the project and then will hold individual consultations with each of the advisory board members over the following six months. In the third phase of the work, the advisory board will review the document outlining the protocol and provide written feedback to the group.

Updates to (and input from) the broader community of RDA will be provided at plenaries every six months in the form of poster sessions, breakout groups, and Birds of a Feather sessions.

## **5. Adoption Plan**

Specific groups committed to taking up the deliverables of this WG include collaborative research projects on the Platform for Experimental and Collaborative Ethnography (PECE). Two instances of PECE – The Asthma Files (TAF-PECE) and The Disaster-STS Research Network (DSTS-PECE) – will provide initial venues for the implementation of the deliverables proposed here. Both have small but active, cooperative, and growing user communities. TAF-PECE is a collaborative research project that currently has approximately 30 consistent users in geographically distributed locations, along with many more student-researchers, all likely to be working on the platform on a daily basis. DSTS-PECE is an international research network that will be actively enrolling new members over the next twelve months -- in groups of five to ten researchers; this incremental enrollment of new members will provide excellent opportunities to test and improve new, embedded metadata management policies. Embedded metadata policies will be part of the PECE Github release and thus will be included in TAF-PECE, DSTS-PECE, and any future instances of the platform. We are aware that the RDA does not promote and endorse specific products and technologies and aim to use PECE as an initial testing ground and then facilitate adoption (implementation of a metadata management plan based on a protocol outlined by the WG) within a number of additional projects.

Other parties interested in this output include:

- Mathers Museum of World Cultures at Indiana University (contact: Jason Baird Jackson, Co-chair of the DPHE-IG)
- Open Folklore (contact: Jason Baird Jackson, Co-chair of the DPHE-IG)
- AstroAnthro.net (contact: Jarita Holbrook, co-chair of this proposed EHM-WG)
- Northwest Knowledge Network (NKN) (contact: Matt Turner)
- National Snow and Ice Data Center (contact: Julia Collins)
- Smithsonian Libraries (contact: Suzanne Pilsk)
- UC Collaboratory for Ethnographic DESIGN (contact: George Marcus)
- Digital Himalaya (contact: Mark Turin)
- Qualitative Data Repository (contact: Dessi Kirilova)
- Perseus Digital Library (contact: Bridget Almas)
- Houston Clean Air Network (contact: Dan Price)
- Berkeley Institute for Data Science (contact: Charlotte Cabasse)
- Philadelphia Health and Environment Ethnography Lab (contact: Alison Kenner)
- DARIAH (contact: Danah Tonne)
- CLARIN
- ICSPR
- Senior scholars that have considerable repositories of research material (currently awaiting our WG deliverables in order to digitize and make shareable their research collections)
  - Michael M.J. Fischer, cultural anthropologist
  - Pat Seed, historian and cartographer
  - Sharon Traweek, cultural anthropologist
- Early-career scholars that are interested in how smart metadata practices might affect their collection of born-digital data, developing informed consent forms, for example, that allow their interlocutors to make a variety of choices about how interviews will be shared.

We will continue to develop relationships with people and projects within and beyond the RDA and aim for a minimum of ten adopters. By connecting to researchers outside of the RDA, a tangential benefit of this WG will be to broaden RDA engagement, especially within the digital humanities.

As the WG concludes its work, we will consider how to expand the protocol to be relevant to other research domains. We will open discussions with members of the Metadata Standards Catalog WG and representatives working on the DMPTool at the California Digital Libraries to consider integrating the protocol in their tools. This may lead to a proposal for a second working group.

Our leadership currently includes representatives from five continents and we are interested in continuing to broaden the geographic (and other) diversity of our membership.

## **6. Initial Membership**

*Leadership (biographic notes in Appendix A)*

- Co-chair: Brandon Costelloe-Kuehn, Rensselaer Polytechnic Institute, Troy, NY, USA
- Co-chair: Dominic DiFranzo, Cornell University, Ithaca, NY, USA
- Co-chair: Jarita C. Holbrook, University of the Western Cape, Cape Town, South Africa
- Co-chair: Lindsay Poirier, Rensselaer Polytechnic Institute, Troy, NY, USA
- Co-chair: Mike Fortun, Rensselaer Polytechnic Institute, Troy, NY, USA

*Initial Members/Interested (based on prior discussions and involvement with the DPHE-IG)*

- Aalok Khandekar, Indian Institute of Technology, Hyderabad, India
- Alison Kenner, Drexel University, Philadelphia, PA, USA
- Anup Kumar Das, Jawaharlal Nehru University, Delhi, India
- Arnost Stanzel, Bayerische Staatsbibliothek, Germany
- Asanobu Kitamoto, National Institute of Informatics, Tokyo, Japan
- Bernard Avril, Strasbourg, Alsace, France
- Brian Callahan, Rensselaer Polytechnic Institute, Troy, NY, USA
- Bridget Almas, Tufts University, Medford, MA, USA
- Clifford Tatum, Leiden University, Leiden, Netherlands
- Dan Price, University of Houston, Houston, TX, USA
- Danah Tonne, Karlsruhe Institute of Technology, DARIAH, Germany
- Dessi Kirilova, Qualitative Data Repository, Syracuse, NY, USA
- Duncan Loxton, University of Technology Sydney, Sydney, Australia
- Ellen Foster, Rensselaer Polytechnic Institute, Troy, NY, USA
- Heidi Lane, University of Helsinki, Helsinki, Finland
- Hilary Goodson, AARNet, Sydney, Australia
- Jamie Wittenberg, Indiana University, Bloomington, IN, USA
- Jason Baird Jackson, Indiana University, Bloomington, IN, USA
- Kai Li, Drexel University, Philadelphia, PA, USA
- Kathleen Fontaine, Rensselaer Polytechnic Institute, Troy, NY, USA
- Kevin Long, Digital Repository of Ireland, Dublin, Ireland
- Kim Fortun, Rensselaer Polytechnic Institute, Troy, NY, USA
- Kimberly Anderson, Iowa State University, IA, USA
- Luis Felipe Rosado Murillo, Berkman Center for Internet and Society, Harvard University
- Lynda Kellam, University of North Carolina at Greensboro University Libraries, Greensboro, NC, USA
- Malcolm Wolski, Griffith University, Brisbane, Australia
- Mark Leggott, Research Data Canada/CANARIE, Ottawa, Canada
- Marta Cywinska, The Warsaw University of Life Sciences, Warsaw, Poland
- Matthew Turner, Northwest Knowledge Network, University of Idaho, ID, USA
- Rainer Stotzka, Karlsruhe Institute of Technology, DARIAH, Germany
- Rebecca Grant, National Library of Ireland, Dublin, Ireland
- Robert R. Downs, Columbia University, New York, NY, USA
- Sara Pittonet Gaiari, Trust-IT Services, Italy
- Sarah Ramdeen, UNC at Chapel Hill SILS, Chapel Hill, NC, USA



- Sharon Traweek, University of California, Los Angeles, CA, USA
- Stephanie Cheviron, Université de Strasbourg, Strasbourg, France
- Tobias Weigel, German Climate Computing Center (DKRZ), Hamburg, Germany
- Vivian Wong, University of California, Los Angeles, CA, USA

## 6. References

Holstein, J.A. and Gubrium, J.F. 2004. "Context: Working it Up, Down and Across," in C. Seale, G. Gobo, J.F. Gubrium and D. Silverman (eds) *Qualitative Research Practice*, London: Sage.

Keller, Evelyn Fox. 2002. *Making Sense of Life: Explaining Biological Development with Models, Metaphors, and Machines*. Cambridge, Massachusetts: Harvard University Press.

Turkle, Sherry, and Seymour Papert. 1990. "Epistemological Pluralism: Styles and Voices within the Computer Culture." *Signs* 16 (1): 128–57.

Riley, Jenn. 2009. "Seeing Standards: A Visualization of the Metadata Universe," available at <http://www.dlib.indiana.edu/~jenlrile/metadatamap/> (Accessed 12/9/2015)

## 7. Appendix A: Leadership Biographical Notes

Brandon Costelloe-Kuehn is an anthropologist and Lecturer in Science & Technology Studies at Rensselaer Polytechnic Institute. Using multi-sited ethnographic methods, his research examines, and participates in, the design of innovative media systems to address the communication and collaboration challenges of politically and scientifically complex environmental issues. He works within a number of collaborative endeavors, including The Asthma Files, PECE and the Multispecies Salon. Brandon was awarded a Summer Internship and then RDA/US Fellowship to to develop the Metadata for Empirical Humanities WG proposal and contribute to a number of ongoing initiatives within the Digital Practices in History and Ethnography Interest Group.

Dr. Dominic DiFranzo is a Post-Doctoral Associate in the Social Media Lab at Cornell University where we works on new design interventions for cyberbullying and the effect of robotics on team collaboration. His research involves collaborating with colleagues across the social sciences and humanities to translate the tools and methods from data science, e-science and informatics to address their research needs and purposes. This includes working with a wide array of research groups and projects including large-scale social network analysis, experimental ethnography, open government data, and web observatories. He holds a PhD in Computer Science from the Rensselaer Polytechnic Institute and was a member of the Tetherless World Constellation.

Jarita C. Holbrook is an Associate Professor of Physics at the University of the Western Cape, South Africa. She holds a doctorate in Astronomy & Astrophysics from the University of California, Santa Cruz. She was a postdoctoral fellow at the Center for the Cultural Studies of Science, Technology, and Medicine at UCLA, and the Max Planck Institute for the History of Science in Berlin, Germany. She is a cultural astronomer focusing on African indigenous astronomy, the culture of astrophysicists and practices of inclusion and exclusion. RDA DPHE-IG members Jarita Holbrook, Sharon Traweek, and Luis Felipe Rosado Murillo are part of AstroAnthro.net, an umbrella project focused on studying astrophysicists, their culture,

diversity and their engagement with big data and big collaborations. Of importance to the group is automating tools for data visualizations characterizing the content of interviews and publicly available data on individual astrophysicists.

Lindsay Poirier is a PhD Candidate in the Science and Technology Studies department at Rensselaer Polytechnic Institute. For the past 3 years, she has served as the lead platform architect for PECE - a role that involves translating the theoretical commitments of the empirical humanities into digital infrastructure. In 2016, Lindsay was awarded a RDA Data Share Fellowship to study the barriers humanists face in adopting infrastructure for data sharing; her work contributed to a number of initiatives in the DPHE-IG. Her dissertation work draws on the history of artificial intelligence and leverages ethnographic methods to analyze the design and politics of the World Wide Web.

Mike Fortun is associate professor in the Department of Science and Technology Studies at Rensselaer Polytechnic Institute. A historian and anthropologist of the life sciences, his current research focuses on the contemporary science, culture, and political economy of genomics. His work in the life sciences has covered the policy, scientific, and social history of the Human Genome Project in the U.S.; the growth of commercial genomics and bioinformatics in the speculative economies of the 1990s; and the emergence of transdisciplinary efforts in toxicogenomics, air quality, and environmental health. Mike Fortun is a co-chair of the DPHE-IG, and a lead developer of PECE.