Prepublication version: please consult with the author (anthlinguist@ku.edu) before citing.

Bringing Digital Data Management into Methods Courses

American Anthropological Association

Linguistic Anthropology module

Arienne M. Dwyer

version 2016-06-10 (revision history)

Author note: The ppt/rtf formats required here by the American Anthropological Association do not meet sustainability requirements for current best data practices.

- **What's wrong with Powerpoint?** Being dependent on a sole piece of commercial software whose source code is secret is a bad idea because it may be unusable to other researchers: you need to buy (and keep buying) the software to read the file. We've all had one or more older electronic documents that we are no longer able to open. It's also an awkward format to get the content out of. As long as one has sustainable formats, having ppt as an *alternate* form is fine.
- **But isn't at least rtf an open format**? No - It's also a closed proprietary format from the Microsoft Corporation, and only word-processing software can open it.
- **What formats are recommended?** For a text-based document like this one, text with markup and a schema (xml, xhtml), even plain ASCII text (txt, with Unicode (UTF-8) encoding); see the Library of Congress's Sustainable format guidelines. If word-processing software is essential (it's not for this module), then software with published (open) XML specs like Open Office is a better and sustainable choice. See Unit 2 "*The Basics*" below for more information.
- **Isn't open format software harder to use and has less functionality?** Not necessarily. Let's not pass on our own biases to other researchers: if someone is just starting out, it will probably be just as easy for them to learn both; using one does not exclude the other.

Aim: To introduce best practices in data management for researchers in linguistic anthropology.
Timeline: This unit can serve either as a one- or two-week standard university course or a short-term (e.g. 1.5 day) intensive workshop.

Prepublication version: please consult with the author (anthlinguist@ku.edu) before citing.

Target audience in any country: (Post-)Graduate students before and after data collection; post-doctoral researchers; early-and mid-career faculty; community members; collaborative projects.

Table of Contents

Course Guide

This course guide may be used as a course introduction.

1. **Data management** is crucial for good research. Scholarship creates a range of data forms that require converting, analyzing, storing, and sharing. As scholars, we have a responsibility to make sure that data endure into the future in accessible formats. These data are gathered by researchers (often from participants), and usually receive input from many people. We are therefore also are responsible for the ethical, legal and intellectual property issues arising from these data, including proper attribution/anonymization and adhering to conventions and laws of relevant locales. Archiving and sharing the output of research in print and online venues (a.k.a. publication) requires attending to best practices in data management. Most research funders now require a data management plan, in order that your results and data be enduring and public.

2. **Beyond scope: This course does *not* cover** methods of obtaining grants or human subjects permission. It is not a tutorial on intellectual property or other national or international laws. It is also not a guide to digitization or format conversion. Some further resources on these topics are found in the references and appendices; some will need to be added.

3. **Data management and archiving begin at *research design***, not after the data are collected.

4. **Ethics begins at research design**: we have a responsibility to plan and carry out research in partnership with a community; to ensure that the work benefits that community, as well as our institutions, our funders, and ourselves; to ensure that our work conforms to local moral and ethical practices, institutional regulations, as well as national and international laws.

5. **Data types in linguistic anthropology**: Linguistic anthropological methods reflect the inter-disciplinary nature of the sub-discipline, and overlap with qualitative and quantitative methods for linguistics (e.g. documentary linguistics, sociolinguistics, discourse analysis, cognitive

tasks) and cultural anthropology (participant observation, interviewing, surveying, and reflexivity). Linguistic anthropologists are likely to work with human subjects. They are likely to generate data that includes any or all of the following: audio and/or video (A/V) recordings and transcriptions of them (often with translation and grammatical annotation, which is sometimes time-aligned); notes, sketches, images (photographs, maps (georectified or not), and diagrams), spatial data, artifacts and other physical data; websites, blogs, emails or other Internet-based communication; ultrasound, and MRI; word lists, grammatical paradigms, sentences, grammaticality judgments of them, and texts; the texts may include printed, handwritten, or electronic texts, questionnaires, surveys, and inscriptions, as well as metadata about these primary research data. These data may be digital or non-digital, structured or unstructured.

6. **Pretty good practice is good enough**. Good practices are not out of reach; don't let "best practices" or this course keep you from learning *pretty good practice* (EMELD 2006).

7. **Key data management practices** are common to all anthropologists. To maximize access to and ethical use of anthropological data, the AAA advocates unified guidelines for data management areas in common among all sub-disciplines. Data management methods common to all anthropologists **can be summarized in four points**:

(1) Data should be put into an **enduring format;**
(2) Data should be discoverable **via metadata;**
(3) Data should be **archived;** and
(4) Data gathering, archiving, and dissemination should be **fully consultative and with the permission of involved participants.**

Course Content

The course sessions present an overview of the key issues in data management via a tip of the iceberg approach: the digital data workflow from research design and project planning to data creation, data management and analysis to preservation, reusability and publication. Each of the five numbered units (four main units and one optional unit) can constitute one or more class session(s); the contents can be covered in more or less detail depending on available time. Each unit requires participants to come up with examples from their own experience and apply that unit's concepts to those examples. Ideally, the instructor (and/or a future version of this course) would also provide use cases for each unit.

1 Ensuring the future of your data and avoiding catastrophe

Aim: To provide an overview of the issues; each bullet is relevant, no matter what the project. Each of these introductory issues is revisited later in the course. Bullet points can be exemplified by the instructor.

- **Workflow: The data lifecycle**
  - What constitutes data? Participants give examples, brainstorm its path through the workflow
  - Data in, data out: input/output formats; planning for output at the time of input
    - *in situ* ("field") work or other research creates input (recordings, notes, maps, images)

        - Analysis and archiving creates output (transcriptions, translations, annotations, lessons, articles, websites, books)
    - Archival, Working, and Presentation data formats (Simons 2006): a useful distinction
        - Archival: highest-quality, lossless (uncompressed), non-proprietary, structured
        - Working: whatever format and size that allows for data analysis
        - Presentation: optimized for sharing (e.g. web/print), so usually compressed audio-visual (A/V) and highly formatted text are preferable

Discussion:

- Is a word processing document (e.g. .docx, .odt) document an Archival, Working, or Presentation format? An mp3 audio? A .tiff image? (N.B. Many A/V formats can be more or less compressed, and for archival purposes, less compressed (less lossy) is always better.)
- Why might an archival format be awkward to work with or share? Why might a presentation format be a poor choice for archiving?

    - **Ethical and legal considerations** (for details see Unit 3, *What are our responsibilities?*)

Planning data management entails negotiation with research subjects and/or a community, about:

        - Informed consent for the collecting, archiving, and sharing of data
        - Access to and availability of data, and
        - Issues of intellectual property.

See Unit 3 and the Appendices below.

    - **Common data disasters** and workflow inefficiencies (exemplified by instructor), e.g.:
        - losing the sole copy of one's data, by failing to back up one's *in situ* research
            - solution: regular backup (LOCKSS)
        - storing data in an obsolete proprietary format, such that it can no longer be read
            - solution: use open formats and consider storing in multiple formats
        - using a non-Unicode font, and later having a "character salad"
            - solution: use a Unicode-based font (there are hundreds, but e.g. Arial Unicode)
        - overwriting a file with an inferior copy with the same name
            - solution: versioning (better) or unique file naming
        - a collaborative team naming files every which way, including *wedding.wav*.
            - solution: systematic file naming

- **Data Management Plans** (DMPs) and their link to the data lifecycle
    - Concerns the sharing and protecting data and participants
    - Motivations: funding requirement; dissertation/publication; better analysis; data reuse.
    - Our responsibilities: sharing *maximally* and *ethically*
        - Maximally, because a greater number of people can (1) re-use the data; (2) learn and benefit from your work; (3) learn about you and your scholarship; (4) benefit from your funder, which helps justify continued funding. (read more: Van den Eynden & Bishop 2014)
        - Ethically, because we (1) enact the wishes of all project participants (especially those of the community, and including data access and credit); (2) attend to local, national,

disciplinary, and funder ethical practices and codes; (3) do no harm, including unintentional harm.

- **What's in a DMP?**
  - Ethics (research design, data gathering, collaboration, sharing, attribution, privacy, etc.)
  - Data (raw and processed): What constitutes the data and its (required accompanying) metadata?
  - Metadata and project documentation
  - Data formats and needed tools: In what formats will the data and relevant tools be archived?
  - Data sustainability and archiving plan: How exactly will the materials be archived and preserved?
  - A named archive: At which archive will the materials be hosted?
  - Principles of access, attribution, and privacy for collaboration and sharing, including:
    - roles and responsibilities of participants: Who does what? How are they credited?
    - production, access, and sharing of research products and their re-use: How and who? And how much? Who controls the data?
    - relevant intellectual property rights (IPR): nation-states' perspective
    - relevant IPR: Indigenous perspectives (e.g. Nichols et al. 2010)
    - local, national and international laws
- Further reading: e.g. Linguistic Data Consortium DMP resources

Discussion:

- Video and notes of Speaker A recorded with consent by Local Researcher B and analyzed by graduate student C, needing to go on Professor D's website. Scenario 1: What will go into the DMP about crediting participants in a research paper? How about when publishing the video online? Scenario 2: How will the DMP attend to changes in consent? (E.g. a consenting participant wants to be anonymized or recognized; a consenting participant withdraws consent; community leaders or a participant stipulate(s) that part of the archived materials be closed to the public).
- How does your own relationship with the community affect your research design?

Exercise: Create a first draft of a Data Management Plan (1-2 pp.) and answer the two reflection questions.

- Use an existing DMP (e.g. NSF, NEH) as a kind of a checklist, to make sure all elements are included.
- Who is your target: what institution's DMP will you use? (E.g. funding agency, First Nations/tribal institution, etc.); for grad students, e.g. NSF-DDIG
- What elements *must* be included? (E.g. permission letter from community; permission letter from archive; acceptable archival formats; research locale and community; research scope; data backup and sustainability; etc.)
- What elements are specific to your project? (e.g. community-specific data access and collaboration requirements; community research product desiderata; political or social

considerations; special data types; special strategies to protect media in a particularly humid or cold climate; etc. Not all of these belong in a DMP, but all will affect research design.)
- Yours will likely include most of the following elements listed in *What's in a DMP?* above.

Reflection questions
- What difficulties (logistical, methodological, or other) did you identify in the process?
- How do the above requirements change your project design, if at all?

2 The basics: Working with data

Aim: To introduce the many data types and formats, and to describe the minimum a researcher must do to create enduring data. Part two (regarding software) will need regular updating.

Unit 2, Part One: Data and metadata

Projects that create digital data during research ("in the field") may need immediate storage for large files (e.g A/V recordings and their associated metadata). DMPs describe each "field" and archival data type, and follow best practices for storing originals and altered versions.

- **Digital and non-digital data**
  - Born-digital data: e.g. "field" recordings, images, geolocations, experimental data
    - Require associated metadata (to understand the scope and organization of the data)
    - May be quite large; require external storage planning (e.g. hard drive or Internet-based)
    - Archiving may require conversion to open formats
    - Require systematic file-naming, organizing, versioning (always retain an original, unchanged version), and backup

  - Non-digital data: print; extant collection or archive; pre-existing dictionaries and grammars
    - May require hundreds of hours to digitize and structure
    - Processes may be different from born-digital data, e.g. :
      - creating a dictionary from a *digital wordlist* only requires structuring and possibly encoding conversion;
      - digitizing a *print dictionary* requires either scanning and OCR or keyboarding, and then structuring.
    - Archival data may also require conversion to open formats
    - Also require systematic file-naming, organizing, versioning (always retain an original, unchanged version), and backup

- **Metadata**

"Metadata is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource." (NISO 2004).
  - Like a bibliographic entry (book citation), but for your data
  - Provides the *necessary context* about your data, allowing access and retrieval

  - Types of metadata: descriptive vs. structural; also administrative, rights management, preservation
  - Which elements, and how much is enough are your decision
  - One possible metadata standard: OLAC 2008

- **Data quality**
  - Data quality: lossless (or less lossy), documented with metadata.
  - Data quality requirements do vary depending on research purpose.
    - E.g. phonetic analysis will generally require higher-quality audio than syntactic work.
  - Still: archive the highest possible data quality, regardless of your topic or discipline.
- **Data capture, regularization, and organization**
  - Digitization, regularization, conventionalization, conversion
    - **Digitization**: representing analog data (sound, activity, object, etc.) in a digital format, i.e. into sets of binary numbers.
    - **Regularization**: systematically replacing irregular forms with regular ones, in order to make certain data comparable. (A version of the non-regularized data is retained.) For example, some transcriptions of conversations might be done in the International Phonetic Alphabet, and others in multiple practical orthographies. To make these usable together, one system (e.g. the orthographic) is regularized to the other (e.g. to IPA).
    - **Conventionalization**: Communities share certain linguistic conventions (e.g. orthographic, disciplinary, etc). For example, it is conventional for discourse analysts to transcribe speech in the language's official orthography, adding conventionalized speech annotation marking (such as the Santa Barbara system). Phoneticians or documentary linguistic anthropologists, by contrast, often conventionally transcribe in a phonetic system (e.g. IPA), and those interested in grammar would add interlinear grammatical annotation, conventionalized by the so-called Leipzig Glossing Rules. Ethnographic convention often encourages the identification of key words. And so on.
    - **Conversion**: transforming from one data format or structure to another (e.g. non-Unicode to Unicode character encoding; from a spreadsheet to comma separated text (csv), etc.)

  - Data standards and conventions
    - Data format: What file format are the data stored in? (Best: open formats)
    - Character format: How are the characters encoded? (Best: Unicode (UTF); second best: (lower) ASCII).
    - File naming - consistent, documented, short, no whitespace or upper ASCII characters.
    - File structure - consistent, documented, avoid over-use of folders

  - Organize data and metadata (spreadsheets, databases, structured text, audio, video, etc.)
  - Document the above activities and conventions
  - Need more info?
    - Try Digital Preservation 101 from POWRR
    - Avoiding problems: cf. problematic file formats case study

- **Well-formed data is enduring data**
  - What is it? Distinguishing the form and content of data
    - Well-formedness: durable and reusable
    - Interoperable - document formats and data structures (e.g. in a readme.txt), so that your data can be opened with a wide range of generic software on any platform
  - Keep data in non-proprietary (open) formats
    - Proprietary (closed) formats: encoding is a trade secret of a commercial company; you generally have to purchase their software (and keep purchasing the latest version) in order to decode the data. If you don't, you lose your content.
    - Non-proprietary (open) formats: published and open to everyone.
    - In between the above two types, there are also (1) "open" proprietary formats: the company publishes the specifications, e.g. mp3; (2) formerly closed formats whose specs are recently published, but whose code is not re-usable, like Microsoft's Open Specification Promise; (3) closed proprietary formats that are not publicly documented, but at least are designed to interoperate in a limited way, e.g. rtf).
  - The same data allow multiple possible outputs if well-formed

- **Discussion exercises** of data and metadata case studies

Instructor: can provide examples of data/metadata in discourse analysis, documentary ethnolinguistics, language socialization

Participants:

- How do you create data? In what formats? How is it organized?
- If you start out a project with the goal of open access (at least some of the primary data must be publicly accessible), how does that choice affect your working with speakers and creating a data collection?
- In particular, how do you respect community norms, which will almost certainly show that not all data can be freely available to the public?
- Name at least one born-digital data type and one non-digital data type that you might use. What issues should your digital data management plan take into account? How might these issues differ for born-digital vs. non-digital data?
- Suppose a Martian (who speaks your language) discovers your miraculously-preserved data in 100 years. What documentation would you need to include, to make sure the Martian can open, and understand your data?

Unit 2, Part Two: Tools (software)

Specific tools rapidly become obsolete; these will need regular updating. In any case, open-source tools that allow maximal re-use are preferable.

- **How tools fits into the workflow**; proprietary vs. open-source tools

  - Transcription tools
    - For discourse and conversation analysis (DA/CA): no special software needed; optionally EXMARaLDA
    - For time-alignment: Praat (originally for phonetic analysis), Transcriber
  - Multi-level annotation tools for A/V (including time alignment): ANVIL, ELAN,
  - Producing Interlinear Glossed Text (minimally: transcription, grammatical glossing, and translation)
    -
  - Digital Humanities tools of use to linguistic anthropologists
    - To start out, web-based interfaces are easy, e.g. text analysis tools e.g. Taporware allow concordancing, frequency counts, etc.; online visualization tools like Vidi to map or graph data.
    - For more functionality and customization, however, locally-installed tools (which may be command-line or GUI) are useful, including R (and R Studio) for both quantitative research and visualization, Gephi for network visualization, and so on.

- At start of project: get the necessary tools that allow you to collect, organize and work with digital data.
- (*optional topic*) Further workflow: corpus development, lexicon, interlinear glossed texts
  - Lexical, text, and other databases

- **Discussion exercises**
  - Discuss knowledge-sharing between participants, of the benefits and limitations of the linguistic anthropology tools you've used.
  - Again, imagine a Martian discovers your well-preserved data. What *tools* might you need to include (if any), in addition to documentation, to allow the Martian to *re-use* your data?

3 What are our responsibilities?

Aim: To discuss researchers' ethical and legal responsibilities and Intellectual Property Rights. It's best to confront these issues during project planning, well *before* an IRB application. Attention to ethics equals good data. Also emphasized are the limits of Open Access: full consultation with communities forms the basis for solid data management plans and sharing arrangements that align with community norms.

- Responsibilities include ethics, rights, and legal issues
  - **Ethical** (and possibly legal) **obligation**: Consent (oral vs. written); attribution of value
    - Informed consent: participation in research is voluntary, and research participants must be fully aware of the purpose of the research, how the data will be gathered, how their privacy (and/or recognition) will be handled, and how the data will be shared.
    - Attribution of value: recognizing the Traditional Knowledge (a.k.a. TEK, Traditional

Ecological Knowledge) of indigenous and local peoples, that is, their holistic collective knowledge on the environment, including physical, social, and spiritual ecosystems. (see e.g. UNESCO 2003, http://www.ser.org/iprn/traditional-ecological-knowledge).

- **Ethical/moral obligation**: sharing of data and research results is *obligatory* in two contexts:
  - For community-based research, respecting community norms for data access
    - The Open Access mantra: "information wants to be free" (Stewart Brand)
    - The community, not just the researcher, determines access (who? how (much)?)
      - project data are co-owned by researcher and community
      - what is shared depends on the types of data and the research context
      - indigenous communities may also have their own protocols (see Legal, below)
  - "Giving back": to the individuals and community in which the data were collected, in a format, in a language, and with content that can be used locally, while attending to privacy concerns of participants.
  - Sharing research results with the public (results and at least some of the data, unless restricted)

- **Legal obligations**
  - University IRB (institutional review board, a.k.a. independent ethics committee (IEC), ethical review board (ERB), research ethics board (REB)): a committee that reviews and monitors projects involving "human subjects."
  - Required at most North American institutions
  - Separately, approval of other IRBs may be needed: e.g. school boards and Tribal IRBs.

- **Moral obligations**: Beyond what the IRB requires, researchers should:
  - Do no (even unintentional) harm; and
  - Arguably, create a research product that is useful to the native-speaker participants.

- Moral obligations sometimes appear at odds with legal obligations, e.g.
  - Requesting *written* permission from "a community" as an IRB requires often sows mistrust (Dwyer 2006, cf. van Driem 2016)
  - Research products useful to communities (e.g. pedagogical materials, children's books) are usually not allowable expenses by funders
  - Such difficulties do not justify avoiding these legal and moral obligations;
  - Resolving these issues is usually community- and project-specific.

- Examples of key conflicts linguistic anthropology (instructor supplies additional examples)
  - E.g. Withdrawing "informed consent" on dictionary making
  - ...

- **Rights**: Intellectual Property Rights (IPR), other rights

- IPR: about ownership of "creations of the mind" (see Newman 2007; Levine 2016)
  - Based on Western notion of ownership, but can include TEK
  - Copyright: who owns and can distribute a particular work (see e.g. UKy Ling 2016)
- Recognition: All participants have the right to receive and credit for their contributions
  - E.g. Co-authorship, and/or citing the specific contributions of all participants in the metadata of research data, as well as research products
- Anonymity: participants have the right to be anonymized in the case of sensitive data.
  - E.g. Using alphanumeric identifiers (rather than names) for speakers, and
  - Ensuring that speakers are not identifiable by public audio-visual materials
  - Not all data can be effectively anonymized

- **Legalities**: Projects subject to the laws of host and researcher country and international law, at a minimum.

- Taking ethics, rights, and laws together, our responsibilities include:
  - **Agreements with stakeholders**
    - "Stakeholders" include: participants; research team; local or national bodies; funding bodies; home institution (see Dwyer 2006).
    - Agreements are proposed during research design and re-visited for potential changes throughout the research
    - Agreements concern many key topics: attribution/anonymization, compensation, responsibilities and division of labor, access to and rights in data and field notes, co-authorship on deliverables including publications, data access, archiving plan and liabilities.
    - Agreements may be written, verbal, or third-party (e.g. via a village leader).
  - International and/or interdisciplinary teams must consider all relevant nations and ethical codes/practices of all relevant disciplines
  - The IPinCH project is an excellent resource, see e.g. Think before you Appropriate.

Discussion:

- Instructor presents examples of legal actions that are ethically dubious and ethical actions that are potentially illegal for the students to debate. Course participants then present examples of how they have/will share data in the two contexts, and discuss how intellectual property rights and legal issues interact with their obligations as researchers.
- Name at least two locally-appropriate steps that can be taken to ensure shared data access by the language community.

Exercise:

- Closed/Limited/Open Access debate: Imagine or enact a role-playing debate between people who take strong positions on the issue of protecting the exploitation of community knowledge vs. "all information wants to be free." Bring up the strongest arguments for each position (with real-life examples), and then see how best a compromise position that addresses all needs is reached. Possible roles (who may be argue any one or multiple sides of the debate): Indigenous

community elders, indigenous linguists, a digital humanist or corpus linguist, an NSF representative, a PhD student (indigenous of the community, indigenous of another community, or non-indigenous), a specialist professor, the university IRB, etc.

4 Archiving and re-use of data

1. **Why should I archive?**

- Many of linguistic anthropology research funders requires archives
- The data (a product of enormous effort) is backed up in a trusted repository
- I and others can re-use the data (subject to any access restrictions)

2. **Data care** (backup and data protection):
    - LOCKSS (Lots of Copies Keeps Stuff Safe); pros and cons of online/offline storage
    - Formats - document your archival, working, and presentation formats (Simons 2006).
    - Versioning - keep track of different versions of the data (possibly with versioning software such as Subversion, a must for collaborative projects)

3. **Key archives** for linguistic anthropologists

- The list (see Appendices) is alarmingly short
- Use the "How to Deposit" guidelines of existing archives to learn more about common data formats, ethical protocols, and best practices

4. **Archival Concepts**:

- Users and use case scenarios
- Access: Fully open, Graded/Tiered (e.g.), Closed (based on confidentiality agreements)
- Ownership: Intellectual property, Copyright
- Deposit: guidelines for which language, data and metadata formats
- Original ("raw") vs. edited data

5. **Mobilization: Re-using your outputs**
    - publishing articles while writing a dissertation
    - publishing a monograph after the dissertation
    - sharing primary data and metadata
    - maximizing re-use potential by others

Discussion:

- When might we or our language consultants not want to share data?
- Will wide data-sharing lead to researchers being "scooped," (i.e., having someone publish your intellectual property before you do)?
- Not all data users will be uni-disciplinary linguists or even academics; what steps can be taken to make the data maximally accessible to and interesting for multidisciplinary groups as well as non-academics (e.g. those in public policy, NGOs, unrelated language communities looking for a possible model, and the public)?

- Data collection can be faulty, preservation imperfect; what steps can be taken to mitigate mistakes?

5 Making the most of your data (Optional additional unit)

Below five separate topics are outlined, whose only commonality is that they are beyond introductory. Each topic awaits further development.

1. **Using Regular Expressions** (RegEx) to convert data into new forms

- RegExes are computational shorthand entered usually into a command line interface
- They allow for more powerful transformations than search and replace.
- General RegEx tutorials or sites, and RegEx for linguists tutorials can be used to learn more.

2. **Working collaboratively** at great distance (remote data access; collaboration environments, ...)
   - documentable collaboration practices (e.g. using a project wiki rather than email and attachments for communication)

3. **Making data and websites accessible** to people of all abilities (e.g. colorblind, hearing/sight impaired, multilingual, non-English speaker, elderly etc.), and to people with slow internet connections. See the W3C's Web Accessibility Initiative recommendations.

4. Establishing your own **digital archive** (optional topic if there's interest)
   - Include any of the following more advanced topics: versioning; linked open data; controlled vocabularies; ontologies; data persistence (location and formats accessible into the future)
   - The trusted repositories mandated by funding agencies are at best in very short supply, and there's little funding to create them.
   - Archiving includes not just data, but also metadata and code.
   - If you already have an archive, how can you improve it? Via assessment metrics ISO 16363 and the Trustworthy Repositories Checklist

5. Planning for **data re-use**
   - In teaching
   - In comparative research
   - In research design
   - In re-analysis (with or without supplementary new data)

Exercises:

- RegEx:
  - Hands on data cleaning, e.g. using data from a website
  - Create 2 tables and then try to merge them (shows how you have to create the conditions for easy re-use.)
  - Using RegEx for simple substitutions
  - RegEx lite: many text editors have some common RegEx features at a menu click; these

may already be on your laptop (e.g. Notepad++ (PC), BBedit or TextWrangler (Mac), Aquamacs, etc.) - give these a try to, for example, convert ALL UPPERCASE LETTERS to all lowercase.
    - Can't be bothered with RegEx? Try cutting and pasting your text into Data Wrangler
- Accessibility: Look over your own or the AAA's website and make accessibility recommendations based on the W3C's guidelines. (2) For your current research project, what are three things you could do to present the data in a more accessible fashion?
- Digital archiving:
    - Look at the ingesting (data-acquisition) requirements for the following two archives: (1) The Language Archive; (2) AILLA. Will they accept anyone's data from any language? What kind of data and metadata formats do they accept?
    - Imagine at least two use cases for each of the two archives above.
    - Imagine at least two re-use cases for each of the two archives above.

Acknowledgements

This module has benefitted from and incorporated the specific comments of Philip Cash Cash, Jenny Cashman, Fatimah Williams Castro, Sara Gonzales, Candace Greene, Jared Lyle, Christine Mallison, Ricardo Punzalan, Thurka Sangaramoorthy, and Stephanie Simms. Naturally, the current author is responsible for any errors or infelicities, and thanks the AAA and the U.S. N.S.F. for its sponsorship and guidance of this effort.

References

Websites mentioned and/or linked in this document do not necessarily represent the views of the author or the American Anthropological Association. Commercial websites mentioned and/or linked here are intended as examples, and do not represent the endorsement of the author or the AAA.

Dwyer, Arienne M. 2006. Ethics and Practicalities of Cooperative Fieldwork and Analysis. In Gippert, Jost, Mosel, Ulrike and Nicolaus Himmelmann, eds. 2006. *Fundamentals of Language Documentation: A Handbook*. Berlin: Mouton de Gruyter, pp. 31-66. Web preprints. [english] [español] https://kuscholarworks.ku.edu/handle/1808/7058

EMELD [Electronic Metastructure for Endangered Languages Data] 2006. Working Group 1 report on Collecting Primary Texts. (Marianna Di Paolo, Gary Holton,\l
Susan Smith, Arienne Dwyer, Steve Moran, Doug Whalen, Julia Good Fox, and Barbara Need.) Web. http://emeld.org/workshop/2006/wg/wg1-report.rtf

Flanders, Julia and Trevor Muñoz. 2016. An Introduction to Humanities Data Curation. Web. https://guide.dhcuration.org/contents/intro/

IPinCH [Intellectual Property Issues in Cultural Heritage Project]. 2016. Web. http://www.sfu.ca/ipinch/

IPinCH. 2016. Factsheet: Traditional Knowledge. Web. http://www.sfu.ca/ipinch/sites/default/files/resources/fact_sheets/ipinch_tk_factsheet_march2016_final_revised.pdf

IPinCH [Intellectual Property Issues in Cultural Heritage Project]. 2015. Think Before You Appropriate. Things to know and questions to ask in order to avoid misappropriating Indigenous cultural heritage. Simon Fraser University: Vancouver. Web. http://www.sfu.ca/ipinch/sites/default/files/resources/teaching_resources/think_before_you_appropriate_jan_2016.pdf

Levine, Melissa. 2016. Policy, Practice and Law. In Flanders and Muñoz. Web. https://guide.dhcuration.org/contents/policy-practice-and-law/

Library of Congress. 2013. Sustainability of Digital Formats. Web. http://www.digitalpreservation.gov/formats/

Library of Congress. 2013. Recommended formats. Web. http://www.loc.gov/preservation/resources/rfs/index.html

Newman, Paul. 2007. Copyright Essentials for Linguists. *Language Documentation and Conservation* 1.1. Web. http://scholarspace.manoa.hawaii.edu/bitstream/handle/10125/1724/newman.html

Nichols, George, Catherine Bell, Rosemary Coombe, John R. Welch, Brian Noble, Jane Anderson, Kelly Bannister, and Joe Watkins. 2010. Intellectual Property Issues in Heritage Management Part 2: Legal Dimensions, Ethical Considerations, and Collaborative Research Practices. *Cultural Heritage Management* 3.1:117-147.

NISO [National Information Standards Organization] 2004. *Understanding Metadata*. Web. http://niso.org/publications/press/UnderstandingMetadata.pdf

OLAC [Open Languages Archiving Community] 2008. Metadata. Web. http://www.language-archives.org/OLAC/metadata.html

Simons, Gary F. 2006. Ensuring that digital data last: The priority of archival form over working form and presentation form. *SIL Electronic Working Papers* 2006-003. Web. http://www-01.sil.org/silewp/2006/003/SILEWP2006-003.htm.

Stanford University Libraries. Best Practices for File Formats. Web. https://library.stanford.edu/research/data-management-services/data-best-practices/best-practices-file-formats

UNESCO. 2003. Best practices on Indigenous Knowledge. Management of Social Transformations Programme. Web. http://www.unesco.org/most/bpikpub.htm

Van den Eynden, V., and L. Bishop. 2014. Incentives and motivations for sharing research data, a researcher's perspective. A Knowledge Exchange Report. Web. http://repository.jisc.ac.uk/5662/1/KE_report-incentives-for-sharing-researchdata.pdf

van Driem, George. 2016. Endangered Language Research and the Moral Depravity of Ethics Protocols. *Language Documentation and Conservation* 10:243-252. [pdf]

W3C [World Wide Web Consortium]. 2016. Data on the Web Best Practices. Latest published version. Web. http://www.w3.org/TR/dwbp/

W3C [World Wide Web Consortium]. 2016. Web Accessibility Initiative. Web. http://www.w3.org/WAI/

Appendices

General Resources for all anthropologists

- Standards
  - Character encoding: Unicode: http://unicode.org

- Data management and DMP planning tools
  - UK data service: https://www.ukdataservice.ac.uk/manage-data
  - DMP online tool: https://dmponline.dcc.ac.uk
  - ICPSR archive data preparation guide [html]

- Data management requirements and sample DMPs from funders
  - U.S. National Science Foundation (NSF)
  - U.S. National Endowment for the Humanities (NEH)

- Data management guidance from institutions
  - Australian National Data Service on Data Management [html] and DMPs [html]
  - Economic and Social Research Council - DMP guidance [pdf]
  - SOAS Endangered Languages Archive - depositing guidelines [html]
  - Linguistic Data Consortium DMP resources [html]
  - ICSPR (Inter-university Consortium for Political and Social Research) Guidelines for effective DMPs [html] [pdf]

- Best practices
  - Library of Congress: http://www.digitalpreservation.gov/formats/
  - see also LOCKSS (Lots of Copies Keeps Stuff Safe) - originated in a 1998 discussion; now is a Stanford website on digital preservation: https://www.lockss.org
  - Open access resources and repositories: http://www.science.gc.ca/default.asp?lang=En&n=ECEFDFAA-1

- Resources on ethics and rights
  - IPinCH [Intellectual Property Issues in Cultural Heritage Project]. 2016. Web.
  - Canadian Institutes of Health Research, Natural Sciences and Engineering Research Council of Canada, and the Social Sciences and Humanities Research Council of Canada. 2014. Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans. [html] [pdf] http://www.pre.ethics.gc.ca/eng/policy-politique/initiatives/tcps2-eptc2/Default/
  - Linguistic Society of America. 2009. Ethics Statement. Web. http://www.linguisticsociety.org/sites/default/files/Ethics_Statement.pdf
  - American Sociological Association. 1999. Code of Ethics. Web. [html] [pdf]
  - American Anthropological Association. 2012. Statement on Ethics. [html]
  - World Archaeological Congress. 1990-2016. Code of Ethics. [html] Statements on indigenous communities and cultural heritage, human remains, and research in conflict zones.

Resources specific to linguistic anthropologists

- Archives and trusted repositories for linguistic anthropology data: Take a look at the OLAC participating archives; some examples include:
  - AILLA: Archive of the Indigenous Languages of Latin America
  - ANLA: Alaska Native Language Archive
  - ELAR: Endangered Language Archive at SOAS: http://elar.soas.ac.uk
  - The Language Archive: https://tla.mpi.nl/
  - PARADISEC: Pacific and Regional Archive for Digital Sources in Endangered Cultures
  - (more to be crowdsourced)

Revision history:

- **draft v.1**, 2016-01-13 (contained the original draft of "Practices common to all disciplines of Anthropology", which forms the bulk of the introduction to all four modules on the AAA website);
- **draft v.2**: 2016-03-13, re-formatted and separated generic data management information; incorporated most of the Course Introduction - drastically shortened below - into a general introduction for all four modules; incorporated workshop #1 expert feedback;
- **draft v.3**: 2016-05-30 (incorporated workshop #2 expert feedback, including adding an ethics section (expanded exercises)
- **draft v.3.5**: 2016-06-10 (expanded; formatted into AAA's requested format)

Feedback on this document is welcome.
Optimization of this module will require regular updating.