LIS 711
Data Management for Information Professionals

Information School
University of Wisconsin-Madison
Fall 2021

Instructor: Dorothea Salo (please call me "Dorothea") salo@wisc.edu, 4261 Helen C. White Hall
Student hours: M 2-4 pm Canvas: https://canvas.wisc.edu/courses/272392
Instructional mode: In person, M 5:30-8 pm Pinboard URL: https://pinboard.in/u:dsalo/t:711

Introduction

Course description

Students completing this course will earn three credit hours. One credit is the learning that takes place in at least 45 hours of learning activities, which include time in lectures or class meetings, in person or online, labs, exams, presentations, tutorials, reading, writing, studying, preparation for any of these activities, and any other learning activities.

This course has no prerequisites or co-requisites.

This course surveys management of many forms of digital data in non-profit, government, research, educational, and industry contexts. It examines organizational requirements and drivers, ethics and legal requirements, design patterns and architecture, infrastructure, processes, human factors, security, and assessment.

Course objectives

- Understand forms, formats, and lifecycles of digital data across a wide breadth of contexts.
- Understand and design for common organizational drivers and requirements for data management practices
- Evaluate software, hardware tools, and sources of reference data relevant across the data lifecycle
- Construct a current-awareness strategy; assimilate substantial amounts of relevant writing
- Self-sufficiently acquire technical and discipline/sector-specific knowledge

MA/LIS: This course is designed to assess student progress in the following program-level learning outcomes: 1, 2, 5.

Course Policies

I aim to make this course as accessible as possible to all students. Students seeking accommodations for lecture or assignments must obtain a McBurney Center Faculty Notification Letter. For more information, see https://mcburney.wisc.edu/apply-for-accommodations/**.**

Name and pronouns: Your name or gender may have been reported to me incorrectly. Please let me know your pronouns and preferred given name or nickname as you are comfortable. My pronouns are she/her/hers. UW-Madison lets students indicate a preferred name: https://registrar.wisc.edu/preferred_name.htm Canvas does as well, adding pronoun specification: https://kb.wisc.edu/luwmad/page.php?id=108069

Contacting me

READ THE SYLLABUS before asking a question, please; the syllabus may answer it! For any difficulty with the course that is not private or confidential, please ask in class; I *will not answer such questions by email*. Please also do your best to assist your classmates.

Should you see dead links (it does happen, usually with no notice), weird due dates, or other syllabus problems, please bring them up in class.

Textbooks

REQUIRED: DAMA, *Data Management Body of Knowledge*, Technics Publications, 2nd edition, 2017 (English-language ISBNs 9781634622349 print, 9781634622363 electronic). Available at https://technicspub.com/dmbok/. You are welcome to purchase an edition in your preferred language if one is available; if you do, you *do not* also need to purchase an English edition. Either print or electronic is fine with me.

Notes from a pandemic

These are not usual times, I'm acutely aware. I am absolutely willing to accommodate sudden unforeseen challenges. Please let me know what you need as soon as you can.

I will broadcast and record in-person classes via Zoom. Anyone may attend class via Zoom for any reason; you need not disclose your reason to me. Please be prepared to participate in in-class activities in Zoom breakout rooms, and please be aware that this class *is synchronous* — watching Zoom recordings afterward does *not* count as class attendance.

If I need to quarantine due to COVID (and I assure you I am fully vaccinated and trying not to become infected), I will do my best to meet class synchronously via Zoom. If I can't manage that, my fallback will be asynchronous-online work. Please help me avoid either of these undesirable scenarios by masking during class and/or attending remotely. Thank you.

Assignments

Grading scale

A 94-100 Outstanding work. Student performance demonstrates full command of course materials. Work shows a degree of synthesis and creativity that surpasses course expectations.
AB 88-93 Very good work. Student performance demonstrates thorough knowledge of course materials. Work shows a degree of synthesis and creativity that is superior.
B 82-87 Good work. Student performance demonstrates the ability to meet designated course expectations. Overall work is at an acceptable level.
BC 77-81 Marginal work. Student performance demonstrates incomplete understanding of course materials. Or student fails to meet deadlines.
C 72-76 Unsatisfactory work. Student performance demonstrates inadequate understanding of course materials. Or student fails to meet deadlines.
D 67-71 Very unsatisfactory work. Student performance demonstrates inadequate understanding of course materials. Or student fails to meet deadlines.
F 66 and below Completely unsatisfactory work. Student performance demonstrates very inadequate understanding of course materials and serious lack of competence on site. Or student misses many deadlines.

Due dates

Due dates below are specified by module (mostly for my reference). Specific due dates/times are in the Canvas calendar.

Assignment	Final-grade %	Due (actual due date in Canvas)
Collective glossary entries	10%	Various
Collective study-guide material	20%	Various
In-class activities	10%	Various
Semester project: three unit reports	30% (10% each)	End of modules 3, 9, 12
Semester project: final report	20%	End of course

Late assignments will be penalized one final-grade percentage point per day or fraction thereof late. I will allow revision and resubmission at my sole discretion and on my schedule only; any student resistance will remove the opportunity.

Collective glossary entries

You will be assigned two course weeks to work on. For each of these weeks, add *at least five* vocabulary items from the assigned DAMA DMBOK chapter(s) to the collective glossary. (I'll make public who is responsible for which weeks on Canvas — work with one another to share the load!) Please sign each glossary entry with your initials in parentheses at the end, so I know to credit you. (I'll resolve any initials clashes the first day of class.)

For each vocabulary item you add, check its definition with other *reputable* sources. If you see a discrepancy, give both definitions and indicate that the incorrect one is from DAMA DMBOK. (Sharp-eyed students: yes, this counts as a DAMA DMBOK inaccuracy that you may write up separately for an extra-credit final grade point!)

This must be in English, please. To the best of my knowledge, DAMA and its registered certification providers do not offer certification tests in any other language.

Collective study-guide material

You will be assigned two course weeks to work on (different from your glossary-entry weeks). For each of these weeks, choose an assigned DAMA DMBOK chapter to summarize *as briefly as possible* in the collective course study guide. Please focus on what you and your classmates most need to know to pass a DAMA certification test! Your summary may be narrative text or a

bullet-point outline, whichever you prefer; if you can formulate your summary (in whole or in part) as a chart, graph, or other visualization, this is welcome also as long as it communicates more clearly than the awful ones in DAMA DMBOK.

I encourage you to approach summarization work with the following questions in mind:

- What is the **problem** being solved here?
- What is the **solution** given in the chapter?
- Why/when/for whom is this the **optimal** solution? What other possible solutions are there, and when (if at all) should they be preferred?

If another student shares a course week with you, please coordinate with them for maximum chapter coverage. I want the end result here to be useful to all of you! As with the glossary entries, please sign the chapter heading for the chapter you work on with your initials in parentheses at the end, so I know to credit you.

This must be in English, please. To the best of my knowledge, DAMA and its registered certification providers do not offer certification tests in any other language.

SEMESTER PROJECT: Tracking campus data

I don't have the computing infrastructure to give you much hands-on enterprise systems experience, unfortunately. What I *can* do, though, is give you a real-life enterprise's data practices, processes, and people to examine: UW-Madison itself!

For each (of three) units, you will turn in a report that applies class readings, discussions, and insights to the data type/source you have chosen to work with. Each class week will list question prompts for this report; address those as they are relevant, but you need not limit yourself to them! Excellent reports will discuss what you have learned and found beyond my prompts.

Your final report, drawing on your earlier reports, will be a *maturity evaluation* of UW-Madison's handling of the data you chose. You may use any appropriate maturity model — that is, you are not limited to the ones in DAMA DMBOK — but you must fully cite whichever one(s) you use. Where you find less-than-full maturity, make grounded suggestions about how UW-Madison can improve.

All students may choose one of the following data types/sources:

- Admissions data (both aggregate and individual; identified, deidentified, and reidentified; any source, including — especially! — web tracking)
- Research data for an established research center or long-running research project
- Canvas/Unizin data (focusing mainly on data about students)

MA/LIS students (only) may also choose one of the following:

- ILS (Alma) data (focusing on patron and usage data, but including collections data)
- UW Digital Collections data

MS/Info students (only) may also choose one of the following:

- Student/Faculty Center data
- Grants data (WISDM; recommended only if you actually have access to it)

If there is a different type/source of UW-Madison data you would like to work with, let me know by the second week of class. There are lots of excellent possibilities I haven't listed! You are also welcome to propose any data type/source where you have insider knowledge (e.g. a research project that you work on), though a project where you do not have insider knowledge will certainly teach you a lot about documentation (often through omission).

You are welcome to share useful information you find with other students working on the same or a similar type/source of data; this is not cheating, but workplace-style collaboration! You must write your reports on your own, however, so that I can gauge your ability to interpret what you have found, assess it for good practice, and apply concepts learned in class.

Extra credit

The Data Management Association's *Data Management Body of Knowledge* book (DAMA DMBOK) is a hot mess containing many inaccuracies, some trivial, some extremely serious. Each inaccuracy you locate and report to me in Canvas *before the chapter in question is discussed in class*, along with a quotation from a reputable source correcting the inaccuracy, is worth one final-grade point, up to a maximum of three. To receive an extra-credit point:

- Describe it in the collective class error-reporting GDoc (see Canvas). Add your initials in parentheses at the end.
- Include page and section number of the DMBOK sentence(s) containing the inaccuracy.

- If you are using the English DMBOK, quote the inaccuracy directly. If you are using a non-English version of DMBOK, you may either find the inaccuracy in the English version OR give me a translation to English of the inaccuracy (rough is fine).
- Cite or link to your source correcting the inaccuracy, and quote or explain the correction. (I don't care about citation styles; whichever you prefer is fine.)

I plan to send the collective class list of errors, along with anything you didn't find that I did, to DAMA at semester end.

Why am I requiring you to buy and learn from a hot mess of a book? Two reasons: it's the foundation of the various DAMA certifications, and reading it for errors should be an excellent exercise in critical thinking. I am also trying to limit the damage by offering overlapping and supplementary readings that are *not* a hot mess.

In-class and homework activities

You are responsible for active participation in in-class activities, and for helping create any products or outcomes thereof. I start you all with full credit here; you will lose credit should I discover you are not holding up your end of things. (Yes, this means you don't have to succeed at solving all the problems I pose… but you do have to *give it an honest try*.)

Several classwork/homework activities will be small-scale explorations of processes (e.g. data cleaning, data transformation) that can happen at much larger scales. I believe that witnessing, understanding, and solving problems at a small scale teaches many skills you will need to solve them at larger ones!

READING SCHEDULE

Unit 1: Soylent data is people!

(Reference: the movie *Soylent Green*, https://www.imdb.com/title/tt0070723/)

Module 1: Course introduction. What do data-management professionals do?

Topics: "Big data" and its management Research data and its management. "Frameworks" and "models." "Scale" and how it does and doesn't change things. Jobs in data management. Relevant professional organizations: DAMA, SLA, RDAP, the Carpentries. Relevant certifications; studying for a certification test. How to read the DMBOK book; how it is a hot mess; how to be usefully skeptical of it.

For your unit 1 report: Choose a type/source of data to work with. Find out everything you can about the ***technology environment*** *surrounding it: storage (on-premise, vendor-stored, and/or cloud?), software (find documentation if you can! kb.wisc.edu may help), security provisions (including authentication/authorization), analytics (look for dashboards and reports), anything else interesting (e.g. recent technology migrations).*

To read after class (though we will get to much of it in class):

DMBOK chapters 1, 16.

Green. "Every job posting requires way more experience than I have!" https://www.vice.com/en/article/93w4e5/applying-for-jobs-without-experience

Carhart. "Better GIAC testing with pancakes." https://tisiphone.net/2015/08/18/giac-testing/ (Ignore details of GIAC, which are irrelevant to this class — read this for the study system.)

DAMA. "Certification levels and requirements" https://cdmp.info/about/ and "Exams" https://cdmp.info/exams/

Module 2: Data ethics. Data governance.

Topics: Historical data harms; IBM and the Holocaust; China's ongoing genocide of the Uighur; US oversurveillance of people of color. Privacy and privacy law; GDPR, US state laws. Human-subjects research; IRBs; the ethics of testing without ethical oversight. Bias in Big Data and AI/ML. Data protection impact assessments.

N.b.: we'll discuss technical aspects of privacy protection in our security module.

For your unit 1 report: Describe realistic ethical dilemmas attaching to your data type/source. Find out as much as you can about the regulatory and governance environment surrounding it: federal and state law, UW System and Madison policies, records schedules, any grant/funding-related requirements, responsible individuals and groups.

DMBOK chapters 2 and 3.

"Data ethicist." https://dataingovernment.blog.gov.uk/wp-content/uploads/sites/46/2021/09/Data-Ethicist-DDaT-Capability-Framework-role.pdf

"A framework for ethical decision making." https://www.scu.edu/ethics/ethics-resources/ethical-decision-making/a-framework-for-ethical-decision-making/

Shan and Dorn. "Engineers' moral responsibility: a Confucian perspective." https://doi.org/10.1007/s11948-019-00093-4

Klosowski. "The state of consumer data privacy laws in the US." https://www.nytimes.com/wirecutter/blog/state-of-privacy-laws-in-us/

Dorwart, Zanfir-Fortuna, Girot. "China's new comprehensive data protection law: context, stated objectives, key provisions." https://fpf.org/blog/chinas-new-comprehensive-data-protection-law-context-stated-objectives-key-provisions/ (N.b. I do not read Chinese — if you do and can correct or add nuance to this, I would appreciate it!)

McDermott and Hatemi. "Ethics in field experimentation." https://doi.org/10.1073/pnas.2012021117

"Data protection impact assessments." https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/accountability-and-governance/data-protection-impact-assessments/

Barabas. "To build a better future, designers need to start saying 'no.'" https://onezero.medium.com/refusal-a-beginning-that-starts-with-an-end-2b055bfc14be

Content alerts for the next three readings: genocide, oppression, antisemitism, racism. Please take care of yourselves.

Black. "IBM's role in the Holocaust — what the new documents reveal." https://www.huffpost.com/entry/ibm-holocaust_b_1301691

Byler. "Ghost world." https://logicmag.io/china/ghost-world/

Moy. "A taxonomy of police technology's racial inequity problems." https://www.illinoislawreview.org/print/vol-2021-no-1/a-taxonomy-of-police-technologys-racial-inequity-problems/

As appropriate to your career plans, **one or more** of the following ethics codes (if I missed your career plans, let me know):

American Alliance of Museums. "AAM code of ethics for museums." https://www.aam-us.org/programs/ethics-standards-and-professional-practices/code-of-ethics-for-museums/

American Association of University Professors. "Statement on professional ethics." https://www.aaup.org/report/statement-professional-ethics

American Library Association. "Code of ethics." https://www.ala.org/tools/ethics

American Records Management Association. "Code of ethics." https://www.arma.org/page/IGP_Ethics

Association of Computing Machinery. "Code of ethics and professional conduct." https://www.acm.org/code-of-ethics

Society of American Archivists. "SAA core values statement and code of ethics." https://www2.archivists.org/statements/saa-core-values-statement-and-code-of-ethics

User Experience Professionals Association. "UXPA code of professional conduct." https://uxpa.org/uxpa-code-of-professional-conduct/

N.b. data scientists have no code of ethics that I know of. DAMA doesn't either. Isn't that fascinating.

Module 3: Wrangling people.

Topics: Maturity models; gauging where an organization or system is. Requirements gathering; "pain points;" functional vs. non-functional requirements. Design processes; user input and why it is necessary. UX and what happens if it's bad. Change management; bikeshedding and how to avoid it. Communication; jargon.

*For your unit 1 report: Find out as much as you can about who at UW-Madison and/or UW System (individuals and/or groups) models/designs, chooses vendors for, collects, analyzes, works with, shares, sells, maintains, and disposes of the type/source of data you chose. Also state which (groups of) **people** are represented in some way in this data.*

***UNIT 1 REPORT DUE** the end of this module.*

DAMA DMBOK chapters 15, 17

Qin, Crowston, Kirkland. "A capability maturity model for research data management." https://surface.syr.edu/istpub/184/

Mifsud. "Requirements gathering: a step by step approach." https://usabilitygeek.com/requirements-gathering-user-experience-pt1/

For reference: "Requirements gathering form overview." https://www.virgowebdesign.com/images/REQUIREMENTS_GATHERING_FORM.pdf

Berinato. "Putting yourself in the customer's shoes doesn't work." https://hbr.org/2015/03/putting-yourself-in-the-customers-shoes-doesnt-work

Spool. "Fast path to a great UX: increased exposure hours." https://articles.uie.com/user_exposure_hours/

Jain. "12 unsettling lessons learned trying to make healthcare better." https://www.forbes.com/sites/sachinjain/2020/05/23/12-unsettling-lessons-learned-trying-to-make-healthcare-better/ (Every single word is also true of data management.)

Mannheimer. "Using data dictionary creation as the teaching moment for metadata." https://web.archive.org/web/20150906000347/http://connect.clir.org/blogs/sara-mannheimer/2015/05/21/teaching-moment (Lots of jargon in here that we'll get to later in the course. That's actually part of the point of assigning this now!)

Unit 2: Designs (on) data

Module 4: Data quality and interoperability

Topics: "Fit for purpose." Data-quality criteria. Interoperability desiderata; how data quality affects interoperability. Data standards and standards bodies; ISO, IEEE, etc.. ETL and other data-migration processes; data validation post-migration. How Microsoft Excel is the enemy of data quality, and what to do about that. Assessing data quality. Remediating low-quality data.

Unit 2 report: Find out as much as you can about data-quality engineering around your type/source of data. ("There isn't any" may be all the answer you need or can find! This happens!) Locate and summarize available documentation explaining data-quality expectations to data collectors or end-users. Find out what you can about recent-ish (say, last 5-10 years) and possible upcoming data migrations across technologies and/or standards. How did they (or are they forecast to) go? Next, find out as much as you can about systems with which this data type/source interoperates, and applicable data standards.

(Hint for your final maturity-model report: data quality and interoperability are frequently areas needing improvement.)

DAMA DMBOK chapters 8, 13

"The Quartz guide to bad data." https://github.com/Quartz/bad-data-guide (Note the available translations; it's obviously fine to substitute the one in your preferred language.)

Lohr. "For big-data scientists, 'janitor work' is key hurdle to insights." https://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html (Please ponder the classism in this headline, whether it exists in workplaces also, and what it signifies about organizational willingness to accomplish data-quality work.)

Mesibov. "An audit of some processing effects in aggregated occurrence records." https://zookeys.pensoft.net/articles.php?id=24791 (For anyone who thinks "the computer will fix the data all by itself!")

Broman and Woo. "Data organization in spreadsheets." https://doi.org/10.1080/00031305.2017.1375989

Goldfedder. "Choosing an ETL tool" and "A sample ETL project." In *Building a Data Integration Team*, available from UW-Madison Libraries via https://search.library.wisc.edu/catalog/9913040504202121 (Skim-only for the tool descriptions.)

Module 5: Data architecture

Topics: Business, data, application, and technology architectures. Zachman Framework, TOGAF, FEA. Waterfall, incremental, agile project management; project charters; service-level agreements and MOUs. System and data inventories. Lifecycles and lifecycle diagrams. [Research] data-management plans; replication; tools and boilerplate. Metrics and assessment. (N.b. this DAMA DMBOK chapter is amazingly useless, so I am using this module as a bit of a grab-bag for other useful-to-know things.)

Unit 2 report: Fill out a Zachman Framework grid (please use a current Zachman formulation, not an older one) for your type/source of data, using what you learned in Module 3 to put (individual or group) names to the various perspectives as often as possible. Create EITHER a conceptual model, data-flow diagram, or data-lifecycle diagram for the data, based on what you learned in prior modules.

(Note any holes in your Zachman Framework grid for your final maturity-model report. They may represent holes in the architecture and/or holes in its documentation!)

DAMA DMBOK chapter 4

For reference: "Project management templates." https://drive.google.com/drive/folders/0BwSO-ACc9obXSUI5dlpwTFUxRG8?resourcekey=0-EMdKztebpsq_mWtlTaMoQg

Tupper. "Enterprise architecture frameworks and methodologies." In *Data Architecture: from zen to reality*, available from UW-Madison Libraries via https://search.library.wisc.edu/catalog/9911118324702121

Briney, Coates, Goben. "Foundational practices of research data management." https://doi.org/10.3897/rio.6.e56508

Goben. "The NIH data management and sharing policy: first thoughts." https://hedgehoglibrarian.com/2020/10/30/the-nih-data-management-and-sharing-policy-first-thoughts/

Ball. "Review of data management lifecycle models." https://researchportal.bath.ac.uk/en/publications/review-of-data-management-lifecycle-models (Skim the actual models, looking for commonalities; read the Maturity Model carefully. I share Ball's skepticism about how useful prescriptive and over-idealized models generally are.)

Goldminz. "Ending the tyranny of the measurable." https://medium.com/the-ready/ending-the-tyranny-of-the-measurable-44aea20e6bd7

Gaede, Thornhill, Henry. "Sustainability for project-based collaborative work." https://static1.squarespace.com/static/531a8b89e4b05d85bc4ff055/t/5cacbf87652dea510636c49b/1554825102932/CNI+Presentation.pdf

Sims and Johnson. "Scrum: a breathtakingly brief and agile introduction." https://agilelearninglabs.com/resources/scrum-introduction/

Denning. "Why do managers hate Agile?" https://www.forbes.com/sites/stevedenning/2015/01/26/why-do-managers-hate-agile/?sh=6bf0843a3a57 (Please discount the ugly self-congratulation in this piece.)

Module 6: Data modeling 1: The classic relational database model; SQL; NoSQL

Topics: Database desiderata; ACID; Brewer's theorem. Entities, attributes, tables, keys, joins. ERDs/EERs; data dictionaries. Normalization and denormalization; effects on query construction, response speed. NoSQL and why it exists.

Unit 2 report: Identify any databases (relational or NoSQL) that are part of your data type/source's platform. Find and summarize the major entities in any documentation you can for it (ERD, UML, data dictionary, whatever there is).

DAMA DMBOK chapter 5 (though you may wish to delay reading it until after class; I support that decision!)

Voss. "Databases: how they work, and a brief history." https://seldo.com/posts/databases_how_they_work_and_a_brief_history (Content alert: the occasional f-bomb.)

Peterson. "What is normalization in DBMS (SQL)?" https://www.guru99.com/database-normalization.html (Through 3NF Rules — ignore the rest.)

Vaas. "To SQL or NoSQL? That's the database question." https://arstechnica.com/information-technology/2016/03/to-sql-or-nosql-thats-the-database-question/

Dybka. "Crow's foot notation" https://www.vertabelo.com/blog/crow-s-foot-notation/ "Chen notation" https://www.vertabelo.com/blog/chen-erd-notation/ and (These are the ones I most commonly see. Your mileage may vary! LIS 751 uses crow's-foot.)

Briney. "Data dictionaries." http://dataabinitio.com/?p=454

Module 7: Data modeling 2: Star schema / dimensional modeling, UML

Topics: Data warehouses and why they exist. Star schema / dimensional modeling; dimensions and facts. Data governance, "insider threat," and its interaction with data warehouses; Alma Analytics as a Very Bad Example You Should Avoid.

Unit 2 report: Is there a data warehouse (or other data-analysis portal) available to end-users of your data type/source? If so, summarize its dimensions and facts. Look for other analyses relying on this data (portals, research publications, reports, dashboards, whatever there is) and tie them as best you can to entities, attributes, dimensions, and/or facts you know are present in the data.

DAMA DMBOK chapter 5

Kimball. "Dimensional modeling primer," "Retail sales" and "Education." In *The Data Warehouse Toolkit*, available from UW-Madison Libraries via https://search.library.wisc.edu/catalog/9911031478002121

Dybka. "UML notation" https://www.vertabelo.com/blog/uml-notation/

"Alma Analytics: General FAQs." https://knowledge.exlibrisgroup.com/Alma/Product_Materials/050Alma_FAQs/Analytics/01_General (Ask yourself whether SQL querying and report *creation* privileges are role-based.)

Module 8: Metadata (by and/or for computers and sometimes human beings)

Topics: Defining metadata. Distinguishing it from documentation (which DAMA DMBOK conspicuously fails to do). Statements, records, documents; "structured" and "unstructured" (meta)data. Common types of metadata (business, technical, operational). Collecting and managing metadata; digital asset management and metadata. Controlled vocabularies; taxonomies; ontologies; reference metadata. Tools for working with metadata. Master metadata and why it should be renamed, ugh. Dates in metadata; ISO 8601, the One True and Honest Date Format. Metadata interoperability; RDF, schema.org.

Unit 3 report: Try to identify standards/guidelines (industry or local) from Gilliland's typology (Table 1) in use by your data type/source. ("Master data" is typically an example of a data value standard…) Are there any they could/should be using that they aren't? Speculate as best you can about why not.

DAMA DMBOK chapters 10, 12.

Gilliland. "Setting the stage [for metadata]." https://www.getty.edu/research/publications/electronic_publications/intrometadata/setting.html

Profisee. "Master data management — what, why, how, and who." https://profisee.com/master-data-management-what-why-how-who/

van Hooland and Verborgh. "Linked data." In *Linked data for libraries, archives, and museums* (chapter 2 section 5), available from UW-Madison Libraries at https://search.library.wisc.edu/catalog/9911170142102121

Hellman. "Spoonfeeding library data to search engines." https://go-to-hellman.blogspot.com/2011/07/spoonfeeding-library-data-to-search.html

Munroe. "ISO 8601." https://xkcd.com/1179/ (transcribed and explained at https://www.explainxkcd.com/wiki/index.php/1179:_ISO_8601)

Module 9: Documentation (for human beings)

Topics: README.txt. Documents DO TOO have structure, DAMA DMBOK, and documentors and documentation systems should respect and leverage that; XML, HTML, DITA. Filenaming conventions.

Unit 2 report: Evaluate the human-aimed documentation (all of it!) you have found so far. Is it useful to its intended audience(s)? Complete? Correct? Readable? Accessible (in the "to folks with disabilities" sense; pay especial attention to images)? Well-designed, well-organized, and well-written?

***UNIT 2 REPORT DUE** the end of this module.*

Briney. "README.txt" http://dataabinitio.com/?p=378 "Starting small: file naming conventions" http://dataabinitio.com/?p=14 and "File naming convention worksheet" https://authors.library.caltech.edu/103626/

Gebru et al. "Datasheets for datasets." https://arxiv.org/abs/1803.09010

Hawkins. "Introduction to XML for text." http://www.ultraslavonic.info/intro-to-xml/

Kimber. "What is DITA?" https://www.xml.com/articles/2017/01/19/what-dita/

Bowen. "RTFM? How to write a manual worth reading." https://opensource.com/business/15/5/write-better-docs

Losh. "Teach, don't tell." http://stevelosh.com/blog/2013/09/teach-dont-tell/

For reference: "The hitchhiker's guide to documentation." https://docs-guide.readthedocs.io/en/latest/

Unit 3: Don't give me any more data! cried Tom Thumb

Module 10: Data storage

Topics: Database hardware and configuration. In-memory databases. Performance tuning; indexes, transactions. What DBAs do. ACID vs. BASE. Specialized database applications; GIS. Backing up databases; archiving databases. Storing, backing up, and archiving data that isn't in databases (which DMBOK inexplicably doesn't cover at all).

Unit 3 report: Find out as much as you can about how your chosen type/source of data is stored, backed up, and (if applicable) archived. Is there a records-management schedule available for it?

DAMA DMBOK chapter 6

Aytas. "Big data storage." In *Designing big data platforms* (chapter 4), available through UW-Madison Libraries at https://search.library.wisc.edu/catalog/9913340373402121

"Indexing." https://dataschool.com/sql-optimization/how-indexing-works/

"SQL transactions tutorial." https://database.guide/sql-transactions-tutorial/

Prater. "How to talk to IT about digital preservation." http://digital.library.wisc.edu/1793/78844

"map school." https://mapschool.io/ (As always, you may read this in any available language.)

Module 11: Data security

Topics: Threat modeling. Common threats to enterprise and research data; ransomware; business-email compromise, phishing. Defaults and why they matter. Data security in the cloud. Data minimization. Horror stories and what to learn from them. Deidentification vs. anonymization (I will die on this hill and I expect you to as well!); reidentification; masking, differential privacy, redaction, and other safeguards for data about people; homomorphic encryption. Basic network security; network segmentation, firewalls, DMZs. Authentication and authorization; multi-factor authentication.

Unit 3 report: For once, this applies to ALL OF UW-MADISON, not just your type/source of data. Find data breaches and other security lapses at UW-Madison that were serious enough to make the news!

DAMA DMBOK chapter 7 (For those seeking errors to write up for extra credit, section 1.3.8 has several serious ones. Bother me in class if I don't remember to correct it, please! There are other errors in this chapter as well.)

Starks. "Cyber insurance market encounters 'crisis moment' as ransomware costs pile up." https://www.cyberscoop.com/cyber-insurance-ransomware-crisis/

Charlton. "Data exchange and the art of iterating security checkups." https://chooseprivacyeveryday.org/data-exchange-and-the-art-of-iterating-security-checkups/

"How security flaws work: SQL injection." https://arstechnica.com/information-technology/2016/10/how-security-flaws-work-sql-injection/

Wodinsky. "Anonymized [sic] data is meaningless bullshit." https://gizmodo.com/anonymized-data-is-meaningless-bullshit-1841429952 (Sigh. Wodinsky means "deidentified." We'll discuss the difference in class, if I haven't already. I agree with the rest of the headline, though, and so do expert privacy and security researchers.)

"The complete guide on data masking." https://www.techfunnel.com/information-technology/data-masking/

Desfontaines. "Why differential privacy is awesome." https://desfontain.es/privacy/differential-privacy-awesomeness.html

Crane. "What is homomorphic encryption?" https://www.thesslstore.com/blog/what-is-homomorphic-encryption/

Salter. "MongoDB's field-level encryption protects private data — even from DBAs." https://arstechnica.com/information-technology/2020/04/mongodbs-field-level-encryption-protects-private-data-even-from-dbas/

Module 12: Document, content, and knowledge management

Topics: Document management; "content" management. Full-text indexing (and its pitfalls); document search/retrieval; relevance ranking. Knowledge management; knowledgebases. Digital asset management. Versioning; git.

Unit 3 report: Take another look at the documentation you found for your type/source of data. What software platform(s) and/or cloud service(s) are being used to manage it? If the platform(s)/service(s) are commercial, try to find their competitors. What's the buzz/zeitgeist/user-satisfaction quotient on those platform(s)/service(s), and why?

***UNIT 3 REPORT DUE** at end of this module*

DAMA DMBOK chapter 9

Flagg. "Content management systems toolkit." https://digital.gov/2013/10/30/content-management-systems-toolkit/

Ferrara. "Strategies for improving enterprise search." https://boxesandarrows.com/strategies-for-improving-enterprise-search/

Middleton. "How the New York Philharmonic became a content management maestro." https://diginomica.com/encore-new-york-philharmonic-became-content-management-maestro

Goodman. "Trial and error: file sharing." https://www.insidehighered.com/digital-learning/article/2017/05/31/trial-and-error-university-arizonas-digital-asset-management

"Guidelines for policy development at UW-Madison." https://development.policy.wisc.edu/ and "FAQs" https://development.policy.wisc.edu/frequently-asked-questions-faqs/ (As an example of enterprise knowledge management.)

Bryan. "Excuse me, do you have a moment to talk about version control?" https://peerj.com/preprints/3159/

Module 13: Big Data and more on data warehouses / datamarts

Topics: Defining "big data." Data lakes vs. data warehouses vs. datamarts. Defining "business intelligence" (paradox or no? sorry, couldn't resist). Assessment. Big Data and data quality. Big Data and data governance. Big Data and data (lack of) ethics; data brokers, data snake-oil salesfolk.

(Note: if we're behind, this week/chapter has been designed to be expendable.)

DAMA DMBOK chapter 11

Gorelik. "Historical perspective" (chapter 2) and "From data ponds/Big Data warehouses to data lakes" (chapter 5). In *The Enterprise Data Lake*, available from UW-Madison libraries at https://search.library.wisc.edu/catalog/9913038167902121

Thomas. "Medicine's machine learning problem." https://bostonreview.net/science-nature/rachel-thomas-medicines-machine-learning-problem (Much of the problem reduces to Big Data data-quality issues.)

Jones-Rooy. "I'm a data scientist who is skeptical about data." https://qz.com/1664575/is-data-science-legit/

Sherman. "Report: Data brokers and sensitive data on US individuals." https://sites.sanford.duke.edu/techpolicy/report-data-brokers-and-sensitive-data-on-u-s-individuals/

Module 14: Managing Big Data and data for data science

Topics: Defining "data science" (and differentiating it from "research that uses data," which is... most research). Data visualization; Tableau. R, Python, SQL, and who might want to learn which to do what; Jupyter notebooks; SciPy, pandas, matplotlib. "Analytics" and various prefixes thereto (including "learning analytics," since we're here). Grid computing, high-throughput computing, high performance computing. "bringing compute to data." In-database computing: MPP. MapReduce.

(Focus REALLY HARD on the "what's solving which problem for whom?" question, or I predict you'll get very lost very fast.)

***MATURITY REPORT DUE** at end of this module*

DAMA DMBOK chapter 14

Gorelik. "Introduction to Big Data and data science" (chapter 3). In *The Enterprise Data Lake*, available from UW-Madison libraries at https://search.library.wisc.edu/catalog/9913038167902121

Chimenti. "High performance computing versus high throughput." https://www.michaelchimenti.com/2015/03/high-performance-computing-versus-high-throughput/

CHTC, UW-Madison. "About our approach [to high-throughput computing]." https://chtc.cs.wisc.edu/approach

Lerman. "Big data and its exclusions." https://www.stanfordlawreview.org/online/privacy-and-big-data-big-data-and-its-exclusions/ (Ask yourself why DAMA DMBOK never discusses determining a dataset's limitations and exclusions.)

"Python vs. R: what's the difference?" https://www.ibm.com/cloud/blog/python-vs-r

MA/LIS learning outcomes

MA/LIS learning outcomes	Course measurable outcomes
1. Students demonstrate understanding of societal, legal, policy, or ethical information issues.	Unit 1 report, especially Module 2's portion. Final maturity report.
2. Students apply principles of information organization.	Unit 2 report, especially Modules 6-8.
5. Students demonstrate competency with information technologies important to the information professions.	In-class activities and homework, especially Modules 4-7. Counted in participation grade.