Agile Data Curation in the Wild: What are your stories?

July 22, 2016

Parallels with Software Development

- What is the problem: the resources that we have for managing data are low, yet the volume of data that is being produced is growing.
    - There has not been an exponential growth of funding.
- Increased demand for documentation, sharing, and preservation. Need more complete documentation. Well acknowledged demand for preservation.
- Moving from a world where researchers were most focused on management. How to organize, structure, format the data. Now there is a much deeper concern of the curation of that data past the initial use of the data.
    - Especially problematic with the exponential growth of data. Need to leverage the growth of efficiencies
    - Need to think about the parallels with software development. How were they designed and executed? How can they be useful in data management and curation? Engineered vs Ad-hoc
- Common processes of software and data curation: process & management tools, agreements & specifications, documentation, product delivery & use.
- On the software development side, there are highly engineered practices. But how do we think critically about how much we invest in those processes? At what point is there diminishing return to the user? Long development cycles.
- A lot of research software and analytics in the Earth Science realm are on the ad-hoc side. We need to come to the middle.
- Agile has attempted to define a middle ground - balancing and counterweighting agility and turnaround time of ad-hoc but also considering the emphasis on well engineered data.
    - Working software over comprehensive documentation
    - Customer collaboration over contract negotiation - need to keep an ongoing conversation in order to manage an ongoing process to ensure you are delivering value.
    - Responding to change over following a plan
- Need to deliver value to the current user and the future user (when we don't know how they will use the data)
- Technical debt: we may accrue debt in the future when data curation is not done in an effective manner. There are certain aspects of your management cycle that you're not investing in sometimes.

Next Steps

- Want to start a larger question of these principles and values with the larger community. Then want to get more concrete and expand our capacity to capture best practices and case studies.

- What we are discussing is not new, but we need to highlight these and spread adoption. "Positive deviance."

Case Studies

- One of the goals: want to get more use cases that are doing agile curation already.
- Scientists have many more opportunities to share their data.
- Have to think of data has having a life story.
- Conceptual framework: along with technical debt, there is data entropy. If you don't do the curation, it loses value over time.

Safeguarding our National Treasure: Nimbus Data

- Want to preserve the data. Lots of different types of tapes and cartridges.
- Have come across some issues: fragile media (from the 1960's)
- Recovery process: request media from the archive centers, conduct independent inventory, ship media to recovery contractor who extracts bits from the tame and provides and image.
- Data recovery issues: bookkeeping, documentation, media, data processing
    - Missing tape labels, missing or extra orbit records, data deviation,s extra end of file and end of tape markers, unknown data formats, duplicate files.
- Used a 64-bit Linux Platform, incorporate C, Perl, shell scripts and ASCII lists.
    - Each tape image is split into individual binary files
- Had to make educated guesses, used trial and error, necessary to have patience and persistence, rely on insight from similar datasets already processed.

Unintentionally Agile: How a Geological Survey Stumbled into Agile Data Management

- Get a lot of information from oil, gas, and goal. They do resource assessment.
- Physical samples and actual objects being curated.
- How to improve data access?
- Legacy data: Working with Data Rescue.
- NGDS: recovery act funding. Started to do data rescue which got them into data curation. Had "digital" but not really. Lacked structure of spreadsheets and multiple versions that people were working in.
    - Whose spreadsheet is better? Not just sifting through and discovering. Have to evaluate it and make connections.
- Don't create a whole new standard sometimes. Individuals and interactions over processes and tools.
- Creating workflows allows for metadata and allows you to discover metadata that you didn't even know you had.
- Archival information science: more product, less process.

- The two case studies are great illustrations of what and why we are doing this. The lessons we can learn from the challenges that they have confronted.

- Incremental value is important: depending on the resources you have available to you. Thinking about what is good enough considering your constraints. Don’t let the prospect of improvement inhibit you from working alongside users today.

Audience Engagement

- Iterative data publication process. Mission side: missions have very fixed data publication dates. But the resources they needed to reach definition of done (in agile terms) weren't coming until the end of the mission. If we had stuck with our model of getting everything complete, couldn't meet that mission completion date. Had to acknowledge tech debt and come back to it.
- But sometimes people will use agile data as an excuse to not do what they need to do. That is what was happening before and why we have all this data that isn’t contextualized.
  - Need to keep current practice in mind in addition to future use
  - Also allows for an entryway. People get stuck in the fact that they don’t want to touch it unless they can do it perfectly.
- Product vs Process: Challenge this to look at this more as prioritization. We only have so much time and funding.
- We have a lot of data to deal with. But need a pragmatic process to identify what data is most important to process.
- A lot of this is more social than science
- Data Centers at NOAA Environmental Information: at the beginning of digital, metadata wasn’t really a process yet. The digital system looked like the physical. Over time lots of data has been collected, but the old data wasn’t curated well.
  - Agile component: needed to get the scientists involved: had hackathons (with food!) to create metadata for these old datasets.
- Management has to buy into the fact that after release is NOT the final stage. You need education, awareness, and socialization with this new way of doing things. Have to be intentional about what you understand as your end state.
- Incremental level of value - data is an ongoing process.
- Responsive Data Curation may be an alternative way to define the underlying values and principles

Conclusion

- How do we start to structure the collection of more and more of these case studies and snippets that have been acquired over the years of data curation processes.
- Those interested should add their case studies via the google link
- There is also a draft surveymonkey on the draft values and principles
  - Especially want to know what we are missing.