Title
Thought pieces for UCLA KI workshop

Author
Thomer, Andrea

Publication Date
2020-02-26

Andrea Thomer
Thought pieces for UCLA KI workshop
9 Jan 2020

1. **What are the most urgent research questions to address about KI? Why?**

Some of the most urgent research questions to address about KI is: *What factors cause, impact, impede, or support KI change? What changes when a KI changes? And when should or shouldn't a KI change?*

Our knowledge infrastructures are meant to be long-lasting – in many cases, meant to last generations – yet, must necessarily be instantiated through materials and organizations that are decidedly less-than-generational in scale and scope: fragile, impermanent technologies like computers, and fluctuating, impermanent social arrangements like "research groups" and "grant-funded projects." Often, KI are maintained over time through piecemeal migration of hardware, software, and data. However, in my work studying database migration in natural history collections (see Thomer et al., 2018) and the maintenance of biological taxonomies (Thomer, Twidale, & Yoder, 2018), I've repeatedly found that we don't yet have a clear picture of the drivers of KI migration. Some seem straightforward, yet are nevertheless challenging to work through, such as hardware and software obsolescence (forced or otherwise), changes in funding streams, or the introduction of new data or data sharing technologies. However, others are less straightforward – or at least, have less straightforward impacts: for instance, organizational shifts as mundane as staff turnover can have widely variable impacts on KI, sometimes leading to a complete overhaul of a system, to the development of extensive work-arounds to keep a system going, or even to the death of a KI. Other more complex sociotechnical changes such as shifts in scientific norms or cultures can have correspondingly complex impacts on a KI. Why do the causes and patterns of KI change matter? Because if we're to build truly long-lived, sustainable infrastructure, we need to account and plan for necessary changes to said infrastructure.

In addition to not having a good model of the reasons underlying KI change, we also don't have a clear understanding of what, exactly, changes when a KI changes – and this lack of clarity can make the management of these vital infrastructures muddled and risky. The computer science literature tends to think of KI migration – or at least database migration – as some combination of changes to data schema, and changes to data format, with some abstract bit of propositional content remaining stable at the center. Yet, there are fields in which true KI migration or updating may entail updates to the data points themselves – for instance, in the field of biological taxonomy, in which scientific consensus on the classification of a given organism may change repeatedly over time. Further, there are social, organizational, and cultural shifts that happen around KI, and that might be viewed as part of a changing KI as well. Further articulation of these closely interrelated aspects of a KI – and their likely differing rates and modes of change over time – will be crucial to supporting infrastructure over time.

2. **Identify a KI whose survival is under threat.**
   a. What led to these threats? Over what time frame?
   b. What actions or changes in circumstances might lead to its survival?
   c. What will be gained or lost, by whom, if this KI fails to survive?
3. **How do KI spread information? Misinformation? Alone and in combination with other infrastructures?**

I bin these questions together because I see the answers as being closely related.

Andrea Thomer
Thought pieces for UCLA KI workshop
9 Jan 2020

I'd like to point to the customized, localized, idiosyncratically-structured collections database as a KI whose survival is under threat, for better or for worse. More and more memory institutions are moving to large, often collaboratively or community-developed collections or data management platforms rather than the ad hoc, idiosyncratic systems they may have begun with during the early days of collections computing. This is perhaps most apparent in natural history museums (NHMs). NHMs were early adopters of database technology, and many developed their own home-grown collections management systems as early as the 1970s. Over the last 20 years, though, these have been gradually migrated to community-developed platforms such as Specify, Arctos, KE EMu, and more.

This move has been largely motivated by a desire to make information management more straightforward (as most NHM database managers have formal training in their domain science, rather than information technology), and to simplify data sharing and aggregation through platforms like the Global Biodiversity Information Facility (GBIF, www.gbif.org) and iDigBio (www.idigbio.org). However, the one-schema-fits-all approach comes with some risks. Not all legacy systems map well to generic schemas offered by these systems; paleontology and other interdisciplinary collections that span domains (e.g. paleo and ecology, geology and biology) have had a notoriously hard time integrating into these databases. A migration to architectures that does not include the entire NHM community can therefore lead to the marginalization of unique, niche or idiosyncratic collections. Additionally, there's a risk that collections might be asked to make "good enough" mappings from their local schemas to these new, shared schemas, which can eventually result in a reduction of data quality.

Thus, the NHM example illustrates a broader issue, in which KI can potentially spread misinformation: through (accidental) misuse of a data schema, and blind trust in a data system as somehow canonical or authoritative. Aggregating information typically requires the use of a shared data and/or metadata schema; sometimes, the migration to this common standard can warp or degrade the information stored. Yet, representation in the aggregate often confers a level of authoritativeness or canonicality to the collection that it would not have otherwise. KI can therefore spread misinformation through the use of the standards they necessarily employ.

References

Thomer, A. K., Twidale, M. B., & Yoder, M. J. (2018). Transforming Taxonomic Interfaces: "Arm's Length" Cooperative Work and the Maintenance of a Long-lived Classification System. *Proc. ACM Hum.-Comput. Interact.*, *2*(CSCW), 173:1–173:23. https://doi.org/10.1145/3274442

Thomer, A. K., Weber, N. M., & Twidale, M. B. (2018). Supporting the long-term maintenance of natural history museum databases. *Proceedings of the Association for Information Science and Technology*.