

Thomas Garnett



Smithsonian Institute, US

GarnettT@si.edu

Thomas Garnett is the associate director for Digital Library and Information Systems for the Smithsonian Institution Libraries. He has coordinated the Biodiversity Heritage Library initiative since its inception in 2004. Recently he was named as the first program director of the Biodiversity Heritage Library. He begins his new position March 31.

Garnett has more than 27 years of experience in the library field, creating, scoping, implementing and managing major digital library projects. For the past 23 years, he has worked in the Smithsonian Institution Libraries, where he served as a system administrator in the Systems Office, before being promoted to assistant director and then associate director for Digital Library and Information Systems.

The Biodiversity Heritage Library is a consortium of 10 natural history, botanical and research institute libraries that collectively hold a substantial amount of the world's published knowledge on biological diversity.

GRL2020 Position Paper

1. Knowledge domains drive research, not libraries.
2. Research data at all levels is produced by scientific projects. It doesn't just pop out of nowhere. Some guided effort has to generate/discover it. Projects can vary from decade-long international ones, e.g. CERN to those involving the "Long Tail of Science" that are managed by an individual.
 - a. Some of this is managed by the projects. Center for Tropical Forest Science <http://www.ctfs.si.edu/doc/datasets.html>
 - b. Some of it, usually more processed and synthesized (cooked), feeds into the publishing supply-chain. <http://www.ctfs.si.edu/doc/publications/articles.html#g> , <http://www.nature.com/nature/journal/vaop/ncurrent/full/nature06557.html>
 - c. Some of it is managed by aggregators
 - i. Some are knowledge-domain specific, e.g. Ocean Biogeographic Information System (OBIS) <http://www.iobis.org/> , National Virtual Observatory (NVO) <http://www.us-vo.org/> These type of aggregators are run by and for scientists and bring value-added service to the data.
 - ii. Some are Institutionally-based – MIT's DSpace <http://dspace.mit.edu/> Often based on services to a broad academic community and a desire to manage the hodge-podge of digital products a heterogeneous faculty produces.
 - iii. Some are more promiscuous, e.g. Internet Archive <http://www.archive.org/index.php>
3. This research data is at multiple levels. Does the "level" determine who should manage it?

- a. Instrumentation data. CO2 levels at estuary locations
 - b. Events – this species of this mosquito at this location at this time bites this marsupial.
 - c. Aggregations – 10 hectare total inventory of every plant in the space
 - d. More examples from biology
 - e. Genomic
 - f. Synthetic – reclassifying an entire arthropod family.
 - g. Multi-domain, complex systems, e.g. climate change.
4. All levels need to be available for the very long-term
 - a. To answer “big” questions in biology requires “big” data.
 - b. Open access is important.
 - c. Interoperability is vital. Silo development efforts are expensive and prevent use of existing data.
 5. The long-term is the “long now” required for actual current use as well. The old-paper-based distinction between texts for current, easy use (libraries) and repositories for primary stuff (archives) will not hold.
 6. Though the data result from knowledge domain-specific projects, the issues associated with curation for the long now are in many aspects not specific to individual research disciplines.
 7. There are great economies of scale possible by sharing the curation of research data while allowing for a certain level of discipline-specific control. What level of control? The economies result from:
 - a. Spreading common costs over a wide base.
 - b. That some problems are not discipline-specific, e.g. issues involved in preserving TIFF files are similar whether the image on the file is a nematode or a galaxy.
 - c. Career pathways for talented staff can be developed, which may be difficult to do with a balkanized approach.
 8. This sharing or aggregation can be centralized or distributed. It may be an organization or a set of practices and documented agreements. It might even be called a “library.” Such a sharing will require an organization that can be virtual or incorporated. The organization will need
 - a. Special skill sets for staff.
 - b. Institutional, scientific, and, national support. The support may include funding, willingness to contribute content, guidance for user-based services.
 - c. Research community buy-in.
 - d. Flexible organizational model.
 - e. Low bureaucratic overhead.
 9. To function it must deal with:
 - a. Political impediments.
 - b. Intellectual property impediments
 - c. Discipline boundary impediments
 - d. Institutional impediments
 - e. Business as usual, inertia
 10. Scientists must have a strong sense of ownership of this aggregation. If it is felt to be distant from ongoing research, they will replicate it for their discipline. But the long-term curation of this will be an insupportable burden on live scientific projects unless common solutions are made available.
 11. The point is not to save research libraries in their current incarnation.
 - a. Instead, the goal is to provide continuity through time to the research record and suite of services to support this.
 - b. To enable new forms of research. Examples include complex systems such as climate change or the relationship between the evolution of the biosphere and the geosphere.