Is data to knowledge as the wasp is to the fig tree? Reconsidering Licklider’s Intergalactic Network in the days of data deluge.

Borgman, Christine L.

Publication Date
2011-07-01

Is data to knowledge as the wasp is to the fig tree?
Reconsidering Licklider's Intergalactic Network in the days of data deluge

Position paper for ICIS Workshop, 23-30 July, 2011:
Accelerating discovery: Human-computer symbiosis 50 years on
https://sites.google.com/site/licklider50/

Christine L. Borgman, Professor & Presidential Chair
Department of Information Studies, University of California, Los Angeles
http://is.gseis.ucla.edu/cborgman/ borgman@gseis.ucla.edu

J.C.R. Licklider famously opened his *Man-Computer Symbiosis* paper [1] with the metaphor of the fig tree and the wasp:

> The fig tree is pollinated only by the insect *Blastophaga grossorun.* The larva of the insect lives in the ovary of the fig tree, and there it gets its food. The tree and the insect are thus heavily interdependent: the tree cannot reproduce without the insect; the insect cannot eat without the tree; together, they constitute not only a viable but a productive and thriving partnership. This cooperative "living together in intimate association, or even close union, of two dissimilar organisms" is called symbiosis.

Lick, in his writings and in his actions, was focused on synergy and joint participation: "Most of the efforts made during the last decade to figure out "what men should do" and "what machines should do" have missed this point widely." [2: 113]. He was particularly interested in information retrieval. His only book was *Libraries of the Future* [2], which laid the foundation for library automation from the 1960s onward, and foresaw many developments in document and text retrieval. Lick advanced Vannevar Bush's Memex ideas [3] from 1945 to 1965, predicting technology for the year 2000:

> We need to substitute for the book a device that will make it easy to transmit information without transporting material, and that will not only present information to people but also process it for them, following procedures they specify, apply, monitor, and, if necessary, revise and reapply. To provide those services, a meld of library and computer is evidently required. [2: 6]

Lick anticipated online catalogs, tablet computers, text mining, and the integration of these technologies into an "Intergalactic Network" [4; 5]. That network was first instantiated as the Arpanet. The Internet Society continues to advance the notion today, with serious assessment of the technologies required to expand networked communication within and beyond our solar system [6].

Some of the challenges that Lick identified were even thornier than he knew, and have yet to be solved. He recognized that interdisciplinary barriers were limiting advances in "pro-cognitive systems," his term for the synergistic knowledge systems that would go beyond Memex. The fields between which "positive interaction" was most required were the library sciences, computer sciences, system sciences, and the behavioral and social sciences [2: 59-60]. These barriers are lower today, and this workshop is an example of positive interaction, but we are far from achieving a deep synergy among these fields of inquiry. Similarly, Lick expected far greater advances in natural language processing, semantic retrieval, and question-answering systems by the year 2000 than has been accomplished yet another decade later.

Given that he was writing about human-computer symbiosis and pro-cognitive systems at the dawn of textual information retrieval [7], it is not surprising that he was focused on relationships among pages (components), books (subsystems), and libraries (systems) [2: 3-4]. The smallest units he considered were words. However, he did pose the provocative question, "Do you suggest that the document read its own print?" [2: 5-6]. Today we do expect documents to be self-describing, at least to the extent of the text *in* the document. The latter constraint is two-fold: (1) Much information necessary to interpret or to validate a document may lie outside its boundaries: Who is responsible for creating this document, what is its provenance, and what is its context? What else must be known to trust this document? (2) Non-textual documents are even less self-describing. Documents containing numbers, images, audio, video and hybrid forms of information pose even greater problems of description and identity. Lick noted the latter problem, at least in passing. His second requirement for "pro-cognitive systems" is that they handle both documents and facts. He defined facts in a footnote:

> "Facts," used here in a broad sense, refers to items of information or knowledge derived from one or more documents and not constrained to the form or forms of the source passages. It refers also to items of information or knowledge in systems or subsystems that do not admit subdivision into documentlike units. [2: 36].

Thus, in some sense, Lick anticipated today's challenges of retrieving data that are not in the form of textual documents. Herein lies another two-part problem: (1) Data are not necessarily derived from documents; more often they precede the creation of documents such as the journal articles that describe them. (2) Data may not exist in the form of static documents; they may be a continuous stream of information, such as observations from sensor networks or telescopes. More problematic is that data are not equivalent to "facts." Facts are assertions that some information is evidence for something.

Back to the fig tree and the wasp: data are knowledge only when humans interpret them. Yet in the days of the data deluge, humans rely heavily on computers for data interpretation. Scholars rely on their instruments and algorithms to clean, verify, visualize, and summarize their data; human eyes may inspect only small portions of datasets. Much can go wrong in the many steps involved in the design and deployment of instruments, collection and cleaning of data, and in the analysis and reporting of results. Data and responsibility pass through many hands, often over the course of many years, in the life cycles of data-driven research. Deep expertise is required in scientific theory, method, instrumentation, and interpretation. Skill sets are complex and are divided differently in each field and specialty. Each step in data handling requires judgment and knowledge of the steps that went before. Necessary details of data provenance often go undocumented, leaving researchers in the position of making multi-party inferences with insufficient information [8]. Minute differences in calibration, miniscule artifacts in a data stream, and other perturbations may be spotted by those closest to the research design – but these factors decrease in visibility the farther the interpreter lies from the source of the data. These risks of misinterpretation multiply as data are combined from multiple sources in the Intergalactic Network and are mined for new interpretations.

The pressure from funding agencies to share research data highlights the complexity of data-driven research: not only the contested notion of "data" itself, but competing views of research, innovation, and scholarship, disparate incentives for collecting and releasing data, the economics and intellectual property of research products, and public policy – and the requisite technical and human infrastructure. The "dirty little secret" behind the promotion of data sharing is that not much sharing may be taking place. Relatively few studies document consistent data release. Sharing research data is thus a conundrum – "an intricate and difficult problem" [9].

Premises:

- Science originates in the synergy of data, computation, and human expertise.
- Data exist in the eye of the beholder; what are data to one person may be metadata, noise, or evidence to others.
- Scientific rewards come from publishing papers, not from publishing data.

Questions:

- What "meld of library and computer" is required to capture, organize, and make data useful across the sciences?
- How can trust be embodied – and embedded – in data systems?
- How can data interpretation transcend the boundaries of disciplines and specialties?
- Where are the social, scientific, and policy incentives to share data on the Intergalactic Network?
- How do we educate wasps to fertilize the next generation of fig trees, and yet assure that they will see the forest and the trees?

RECOMMENDED READING: [9; 1; 2; 5]

REFERENCES:

[1] Licklider, J. C. R. (1960). Man-Computer Symbiosis. *IRE Transactions on Human Factors in Electronics,* **1**: 4-11. Retrieved from http://groups.csail.mit.edu/medg/people/psz/Licklider.html on 5 July 2011.

[2] Licklider, J. C. R. (1965). Libraries of the Future. Cambridge, MA: MIT Press. Retrieved from

http://comminfo.rutgers.edu/~tefko/Courses/e553/Readings/Licklider%20Libraries%20of%20the%20future%201965.pdf on 11 July 2011.

[3] Bush, V. (1945). As we may think. *Atlantic Monthly,* **176**(1): 101-108. Retrieved from http://www.theatlantic.com/doc/194507/bush on 29 December 2008.

[4] Licklider, J. C. R. (1963, April 23). *MEMORANDUM FOR: Members and Affiliates of the Intergalactic Computer Network; SUBJECT: Topics for Discussion at the Forthcoming Meeting*. Advanced Research Projects Agency. Retrieved from http://www.kurzweilai.net/memorandum-for-members-and-affiliates-of-the-intergalactic-computer-network on 11 July 2011.

[5] Waldrop, M. M. (2001). The Dream Machine: J.C.R. Licklider and the Revolution that Made Computing Personal. Cambridge, MA: MIT Press.

[6] *Interplanetary Internet*. (2011). Internet Society. Retrieved from http://www.ipnsig.org/home.htm on 11 July 2011.

[7] Cleverdon, C. W. (1960). The ASLIB Cranfield Research Project on the Comparative Efficiency of Indexing Systems. *Aslib Proceedings,* **12**(12): 421-431.

[8] Meng, X.-L. (2010). Multi-party inference and uncongeniality. In. *International Encyclopedia of Statistical Science*. Berlin, Springer-Verlag.

[9] Borgman, C. L. (2011, submitted). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1869155 on 22 June 2011.