Meditation on Age of (Information) Exploration (Discovery)

Michael A. Keller, Fiesole Conference 2016 remarks at dinner, 7 April 2016 

Dedicated to Mario Casalini, 1926-1998

Please take a few seconds to close your eyes and bring forth images of vast mountain ranges, the great steppes and plains, wide heaving oceans, the endless heavens mapped by our earthbound telescopes and more recently by satellites. Then for a few more seconds, think of the extent of the recorded output of mankind in written form, perhaps 136 million books and maybe a few hundred million printed journal articless, and thousands of miles of archival documents, including the output of governments.

Those of us arising from Western European cultures understand the Age of Exploration or the Age of Discovery to have been the 15th through the 18th centuries during which the New World was discovered, the Earth circumnavigated, and the poles finally visited early in the 20th century. Names such as the Florentine Amerigo Vespucci, one of the quartermasters for Christopher Columbus' voyages and finally a participant in the voyages of discovery from Spain that revealed the full extent of South America in the early 16th century. Vespucci's name was enshrined in the name America by the map maker Martin Waldseemüller. Within 2 decades the Magellan expedition circumnavigated the Earth. All of this is documented. What is not well documented are the voyages of the Vikings to the New World starting apparently in the 10th or 11th century as has recently been emphasized by the discovery of a new site in Newfoundland. Also undocumented are the presumed voyages of St Brendan, one of a series of intrepid Irish explorers and wanderers, who may have arrived in what we now call the Canadian Maritime Provinces in the 13th century. Before the Vikings came to the New World, they were traveling as far away from their homes in Scandinavia starting in the 8th century to Asia, Africa, the Mid East, and of course the British and Irish Isles, Europe, and what is now known as Russia and Ukraine. Very little of this is documented and what we know of it comes from the work of modern archeologists and anthropologists. There is also strong evidence that Chinese exploration occurred over the period 2nd century b.c.e. to 15th century c.e.

And before all of that documented and discoverable exploration, we must confront the dispersion of Homo sapiens from Africa through the rest of the world as a species replacing other hominid forms starting maybe 200,000 years ago. That long period of exploration and really colonization is understood by modern scientists, paleo-anthropologists examining hominid fossils and comparative genomic analysis.

That period of dispersal, let us also call it discovery, lasted, one might say, until the birth of writing in the Middle East roughly 5200 years ago in Mesopotamia and Egypt and the birth of literature about 4200 years ago. Writing and especially written literature confers on human civilization the benefits of recording information consistently for concurrent with the time of the written record and for transmitting information to succeeding generations.

In the Judeo-Christian the importance of the tradition of dealing in writing with the relationships of humans, men and women, with their God through the literatures of the Torah, the Old Testament, the New Testament, the Koran, and the liturgies of the worship of God compelling consistency of approach in theology and narratives mythologizing humankind's relationship to God cannot be overemphasized. Starting with the Dead Sea Scrolls roughly of the 2nd century b.c.e. through the 1st century c.e., the transmission of this tradition continues through Flavius Josephus, then Boethius, and the more or less unknown scribes in the early medieval monasteries and abbeys at places like Reichenau and St. Gall, to the birth of the Renaissance toward the end of the 14th century in Florence. And we all know of the numerous startling changes brought about by Gutenberg's revolution in printing in the mid-15th century.

And that period from roughly the 6th century to the Renaissance is where librarians and publisher arise as intermediaries between authors and readers.

You will notice, I am sure, that this meditation has pivoted from what we might call geographical exploration of relatively ancient times to the amassing of documentation and the transmission of essential texts, themselves giving rise to intellectual exploration and discovery, sometimes thwarted, sometimes flourishing, both functions often occurring simultaneously in various political or theological regimes. Starting with the founding of the University of Bologna in 1088 and the establishment of many medieval universities, that transmission has slowly grown from rote learning of a limited repertory of texts to the present period of phenomenal growth of "discoveries" recorded in hundreds of millions of texts in books and journals and more recently media objects. The honorable tasks of publishers and librarians are to bring the records of discovery and the inventions of creative artists to readers, scholars, and information explorers of the generations alive today and yet to be born.

It is this exploration of information that occupies my thoughts for the rest of this meditation.

Peter Lyman and Hal Varian estimated in "How Much Information", a website amazingly still visible, that in 1999 there was about 2.5 exabytes of information, doubling in 3 years, in 2002, to 5 exabytes. How much is an Exabyte? About one billion gigabytes or about 7.5 times larger than the contents of the Library of Congress. Extrapolating from those estimates to the present time, there would be about 40 exabytes of information, well over 90% of it stored digitally somewhere. Of course the output of publishers plus the Open Access movement plus the data preserved as a result of the requirements for data management plans accounts for only a small fraction of all that information. Much of it resides in the data stored by the astrophysicists, the genomicists, oceanographers, and atmospheric scientists. All sorts of demographic and economic data are held by governments, some of which is accessible to citizens and scholars.

Think of the ways publishers, archivists, and librarians have attempted in the pre-Internet and pre-World Wide Web era to make possible some level of discovery by arrangement on shelves (by size, color of binding, classification schemes) and by various indexing and cataloging systems, often perplexing, frequently not converging with one another. We have created these systems of classification and cataloging at first to account for local holdings, essentially

inventories of investments. In the late 20th century with MARC, OCLC, and RLIN, among others, these evolved to multi-institutional and now multi-national discovery environments for only a part of the information resources available to students, scholars, and readers of any stripe.

Let me illustrate from my own experiences how limited our methodologies to date really are.

In my brief time as a musicologist working on late 16th and early 17th century Italian instrumental music, I confronted indexes by first name, by family name, and by location. I dealt with cataloging "systems" inconsistently dealing with all those elements, sometimes translated from the original language to a temporarily determined common language. Take for example the entry for the journal Acta Musicologica published in Leipzig from 1931 by Breitkopf und Härtel. Acta Musicologica is the journal of the International Musicological Society, based in Basel, bu published in Leipzig originally. It has been and is one of the essential channels of scholarly communication for those in the field of systematic study of music, particularly the history of music. The imprint in the journal itself, intended for an international readership of music historians, showed the place of publication as Lipsiae! So for many catalogs, the only entry for Acta Musicologica was not under title, but under Lipsiae, the Latin form of the German name Leipzig! One had to have honed one's abilities to search for even common, well used titles in many different ways in the pre-Internet era, the vast majority of time in scholarly pursuits in the modern sense from the 19th-century forward. The search for relevant manuscripts, archival documents, and even printed sources was impeded by numerous systems, some of them arcane even at the time. The exchange of letters of potential interest between Italian court officials regarding music and musicians involved tedious and long running searches in bound copies organized by name of secretaries in sending and receiving courts during the lifetimes of the subject musicians and composers.

Clearly the development of a more or less systematically applied set of cataloging rules and the MARC cataloging record across national boundaries in the 1960s and beyond made searching for published sources in OPACs and in the now quite outmoded card catalogs much easier. However, scholars and students desiring to make use of primary sources in the ongoing pursuit of widening the commonwealth of knowledge by professional scholars was only minimally improved by more extensive finding aids, some of which had been published in journals devoted to a single archive, such as Studi e Testi series of the Biblioteca Apostolica Vaticana reporting on holdings of the Vatican library and archive, or in local scholarly society publications. Those improvements have been somewhat superseded by e-journals, provided of course that one understands the need for keyword and key phrase searching in multiple languages and the journals themselves or their e-publishers support that sort of searching. My own search for documents concerning Pietro Lappi, a Florentine monk in the order of Gli Eremiti di San Girolamo da Fiesole, who was known to me as a composer of instrumental music from his final base in the basilica of Santa Maria delle Grazie in the late 16th century. Thanks to an article in Studi e Testi in the 1920s detailing the exchange of collections of archives between the Vatican and the Archivio di Stato of Rome, one could discover that the archive of the order, that of the Eremiti di San Girolamo da Fiesole, an order suppressed in the 1680s in order for the Pope to ransom the city of Candia in Crete from a multi-year siege using the income from the

sale of silver and gold plates and chalices as well as paintings and other art works from several monastic orders. That archive had been taken by Napoleon to the Louvre as booty from his first Italian campaign starting in 1796. Those Italian treasures were returned sometime in the 19th century, after Napoleon's capture and incarceration and through a negotiation between the Vatican and the French government. Alas, those treasures did not go back to the institutions from which they were taken! Thus the importance of the article in Studi e Testi, just to discover the location of the archive finally in the Archivio Segreto Vaticano. Once there, a late 17th century index could be used to identify precisely which archive was of interest and could be called up.

Clearly that was an inefficient and time consuming bit of research that, once successful resulted in weeks of reading archival documents. In this case, the exploration that resulted in the discovery of the possibly relevant documents had its own benefits – shareable serendipitous discoveries, helpful interactions with colleague scholars as well as with the deservedly famous Father Burns, an Irish priest who served readers in the Vatican Archive for a couple of generations, and insights into the vagaries of archival practice in politically active countries and regions.

In short, prior to the World Wide Web, exploration and discovery were challenging tasks, luckily for me requiring direct personal access to Italian libraries and archives, going well beyond the card catalogs of the late 20th century.

The situation for exploration and discovery in the Age of Information, really the initial period of the Internet and the World Wide Web, has definitely improved for e-publications and digitized publications. Keyword and key phrase searching has opened the contents of books and journals that formerly were closed and required lots of reading and note taking, but without accounting for missed possibilities of relevance even in locally held collections due to adherence to limiting rules for description and analysis. The situation in archival collections is not much improved either, though access via email and through e-finding aids, definitely has increased the hints about possibly relevant archives and individual documents from archivists and other scholars.

In the realm of big data, big text collections, and image collections, potential users are dependent upon metadata, some of which may be articles about the relevant data set. Some help seems to be forthcoming, at least in part from the MIT Media Lab. Here is a quote from the NYTimes of Monday of this week, 3 April 2016.

"For years, the federal government, states and some cities have enthusiastically made vast troves of data open to the public. Acres of paper records on demographics, public health, traffic patterns, energy consumption, family incomes and many other topics have been digitized and posted on the web. This abundance of data can be a gold mine for discovery and insights, but finding the nuggets can be arduous, requiring special skills."

A project coming out of the M.I.T. Media Lab on Monday seeks to ease that challenge and to make the value of government data available to a wider audience. The project, called Data USA, bills itself as "the most comprehensive visualization of U.S. public data." It is free, and its software code is open source, meaning that developers can build custom applications by adding other data." From the NYTimes Monday, 3 April 2016.

Particularly regarding keyword and key phrase searching in e-published or digitized sources there are limits. Take, for instance, the now limited results of Google and Bing searches. The extent of possibly relevant sources presented is now limited to perhaps five screens. That situation is not as limited in the case of Google Scholar searches, thank goodness, which shows lots of possibly relevant e-articles from a huge universe of perhaps 200 million, but discerning what is relevant in those large results is an inherent limiter to most searchers, particularly undergraduates.

How can we as publishers, whether primary, secondary, or tertiary, and as librarians and archivists, integrating with our colleagues in computer science, applied mathematics, and our institutional i.t. specialists improve exploration and discovery for ordinary mortals, our readers? Remembering the vast horizons of information described and estimated by Peter Lyman and Hal Varian, how can new methods supersede the rather simple minded approaches of combined methods of metadata creation and presentation involving keyword and key phrase searching? Mention hyperlinking and shared annotations via email and social media…

There are two possibilities to consider, developments that provide some means for navigating vast virtual realms of information: one is linked data leading to the semantic web and the other is algorithmically analyzed contents of e-bases of texts, images, and media for ideas beyond the limits of keywords, perhaps using ontologies and machine generated taxonomies.

Here are a few indicators of activity in this new means of exploration and discovery.

Linked Data – publishers, librarians, archivists creating RDFs and sending them for reconciliation with numerous RDF clouds. Bibframe is a start. Questions of visualization of exploration via the parameters of any idea or any documentary or media object remain and need research as well experience with communities of users. Will most users be content with reconciled notions from the original canon of RDFs and the first couple of halos of RDFs? Will some researchers need and use access to distant halos, out in the distant horizons of reconciled relationships of RDF parameters?

Analyzed contents of e-bases of texts, images, audio and other media files driven either by ontologies and taxonomies or through user interactions with image recognition applications for image-bases or through applications discerning characteristics or patterns or even anomalies in media objects will provide revolutionary possibilities for innovation in exploration. Precursor interfaces and at least one well developed user interface for conceptual exploration of text bases are around; here I refer to the new Yewno conceptual discovery service. These demonstrate that the limitations of keyword searching in full texts and in metadata about texts,

images, and media objects can be overcome. As the costs of CPUs and memory, particularly cloud memories, decline and the efficiency of processing of these e-bases of texts, images, and media objects through streams of applications increases, and as more sophisticated models of user interfaces become more widely usable, I predict the probability of new discoveries of relevance, new hypotheses driving research, especially interdisciplinary research. That these new modes of exploration and discovery will be accessible, even intuitively engaged by secondary school children will in the next couple of decades dramatically improve the rates of discovery, whether for enhancements of our own perceptions and comprehension of wildly complex or even chaotic cultural matters or for re-engineering our approaches to solving vexing problems involving the very survivability of our species and our planet.

Interstellar investigation by way of satellites might be interpreted to address a couple of broad intentions. The first, of course, is the exploration of our universe, voyages of discovery in our day similar to those of Columbus, Vespucci, da Gama, and Magellan, whether for knowledge itself or for commercial purposes. The more consequential reason for interstellar exploration is driven in part by our consumption of the resources of spaceship Earth and by the imaginations of science fiction authors and now film makers starting with Jules Verne. That is, the drive for continuing our species beyond the drive to continue our own selves by raising our children through these engineering marvels starting with the Voyager series and more recently by the Mars Exploration Rovers are meant to discover environments to which we could send human colonizers, hopefully before we collectively exhaust the biomass and possibilities of continually re-constituting that biomass of our Earth.

Of course, one could confront and absorb the recent fantasies of sci-fi driven neuroscientists and computer scientists in providing eternal life eventually to many individual's sensibilities, perceptions, and feelings by copying the contents and constant digestion of ideas and memories in our brains to robotic avatars of ourselves. That is cold comfort, I fear.

The way forward, I think is to accept the grand challenges of linked data and exploitation of both big data and big analysis of big data for a new age of exploration in the present age of information. The Latin phrase "ad astra per aspera", to the stars through difficulties comes to mind.

What, you may ask, should we in this conference support or try to do to move to a truly supportive environment for the age of exploration of the vast & constantly growing array of information enabling new discoveries and new possibilities for understanding? We should all constantly strive to answer that question.

##30##