DATA CURATION INSIGHTS

JOE SEWASH

SERVICES PROGRAM MANAGER

Center for Geographic Information and Analysis

BIG

Big Data Public Private Forum

Introduction

Joe Sewash joined the CGIA in September 2005 and managed the Stream Mapping and Master Address projects. He has over fifteen years of experience in the geospatial field working with public sector clients. Joe contributed to the development of the GIS Certification Institute and has participated in the development of the ISO metadata standard. He is an active member of the National States Geographic Information Council (NSGIC), contributing as board secretary (2000–2005), co-chair of the conference committee (2001–2004), and co-chair of the elections committee (2001–2005). Joe is a graduate of West Virginia University, earning a BA in Geography and Master's Degree in Geography, and currently co-chair of the departmental visiting committee.

Edward Curry: Please tell us a little about yourself and your history in the industry, role at the organisation, past experience?

Joe Sewash: I have an 18 year career in public service. My career has been focused in state level, public sector GIS. Roles have covered tactical products (leading teams in development of multi-use GIS datasets) and strategic development of enterprise architecture and state/national standards.

We use tools from the GIS field for all our work. These technologies haven't been dedicated to the big data side of things. In terms of data preservation people working on the electronic records branch, they've used a combination of tools that have been used by the US Library of Congress from the open source community. One of the areas I have been frustrated with is the fact that you have to work very hard to get the GIS data in these particular programs to work with the tools you would be familiar with on the data analytics side of things. There is a cultural mismatch going back to the days of data warehousing and business intelligence and rolling forward, GIS has traditionally been seen as a mapping component or as a way to visualise traditional BI or other kinds of analysis whereas as a practicing geographer there is an entire suite of analyses that can be spatially enabled and it seems that there is a long way to go to make those things happen.

I am a classically trained geographer but my career has worked into an area of technologist and technology management, understanding that the technologies we are trying to implement ultimately has to meet a business need in order to justify the investment.

I have a BS and masters in geography focusing on GIS, remote sensing and satellite image analysis. My master thesis was on metadata in distributed queries, so there was always this natural mix of both the geography and GIS technology side. Through the three positions that I've held, there has always been public service since we are technically working on consultant form in a public sector context.

> "As the new generation is getting ready for stepping up, and it is already stepping up, it is more tuned to the end-to-end process, implementing the technical solution and demonstrating the results"

Edward Curry: What is the biggest change going on in your industry at this time? Is Big Data playing a role in the change?

Joe Sewash: Two areas relevant to the intersection of Big Data and GIS would be GIS solutions in the cloud, and crowd sourced data. GIS in the cloud has been working through the adoption curve approaching a mature status. Reaching mature status will provide a critical mass of GIS availability for linking to Big Data cloud-based analysis tools. Crowd sourced data source and publicly provided open source GIS datasets provide interesting avenues for considering Big Data and spatial data nexus points.

Crowdsourced data, at least here in America, is commonly denoted as you carry your cell phone, the telephone companies track your distance between the towers and they make inferences that can be commercialised back to show the flow of traffic rates during rush hour. That would be one example of crowdsourced data coming in from mobile platforms. In terms of publicly provided open source data, there are initiatives like Openstreetmaps or open address which provide a website and a thin client that permits users to update datasets that are open source and utilised for multiple purposes. One of the areas that we haven't been involved in as an organization but that I watched as a professional is the issue surrounding data quality and reliability. How do those factors play into the geographical data and how do the users make use of these open datasets? Are they aware of the quality issues or they are just happy to discover and start using it?

Edward Curry: How has your job changed in the past 12 months? How do you expect it to change in the next 3–5 years?

Joe Sewash: US demographics related to retirements of the Baby Boomer generation and transition to the generation known as "Generation X" will lead to a shift in GIS management perspectives. The Baby Boomer generation came of age during a period of the 1980's when spatial data creation preceded analysis; policies, initiatives, and priorities oriented toward a data-centric approach are a hallmark of Baby Boomer management. Generation X managers came of age in the 1990's during the rise of the Internet and a transformational period in GIS development (movement away from

proprietary databases, proprietary scripting languages, and towards a tighter integration with traditional IT). Gen X GIS managers will be more attuned to addressing questions relating to value proposition and deeper integration of GIS in service-based solutions.

The generation that is ready to retire really had to digitise the world. Data in their early career was a blood and sweat proposition and before you could do the geographical analysis in the GIS, you had to literally digitize the geography that you had to study. As these people matured through their carriers and have moved into the leadership roles, their focus and predispositions were towards these kinds of data policies. I'm part of this generation, we saw all these migrations, moving to PC based GIS, moved to commercial technologies for database and the new generation is getting ready for stepping up and it is already stepping up, is more tuned to the end-to-end process, from business problem requirements, implementing the technical solution and demonstrating the results . It is going to be a not so subtle shift in the priorities and way people look at things and the interrelationships between geospatial and Big Data we are really going to see an interesting trend there in the next 15 years.

For the generation Y and millennials people, in particular in geospatial, everything was already built. They are much more focused on thinking creatively and developing solutions with resources that are already available. One of the interesting areas that talked about with several colleagues that are starting families, having children now that will be part of generation Z, from a geographer's concept, the advances in geographic thoughts and the availability of map based technology in tablets, smartphones, really is going to offer opportunities in raising the geographical IQ of the next generation.

Edward Curry: What does data curation mean in your context?

Joe Sewash: CGIA participated in a four year project funded by the US Library of Congress titled the Geospatial Multistate Archive and Preservation Partnership (GeoMAPP). GeoMAPP brought together geospatial and digital preservation practitioners to identify best practices, technical issues, and business planning for preservation of geospatial data. In this context, data curation is addressed through consistent appraisal and preservation workflows.

Edward Curry: What data do you curate? What is the size of dataset?

Joe Sewash: CGIA maintains a geoportal known as NC OneMap (http://www.nconemap.net/) As NC OneMAP datasets are superseded through an update cycle, the superseded datsets are provided to the North Carolina Department of Cultural Resources (NC DCR) for preservation. This process includes both raster and vector geospatial data, so the size of the datasets can vary greatly. Additional detailed information can be provided.

When you visit NC OneMap there are hundreds of datasets, some of which will potentially be superseded and curated. We provide links to data services that other agencies or other groups may provide so OneMap becomes the discovery portal for those. In terms of the vector datasets, those who represent things like street maps, administrative boundary maps, those tend to be very small in size but high in complexity. You might want to have state wide datasets that are smaller than 100MB. In terms of the imagery because of the raster size and because of the compressed and uncompressed volumes, we can be dealing with state wide datasets that are on the order of 20 TB.

End to end, the number of users could be varying dramatically. There are 3–4 people on our side that deal with the database and server maintenance for the NCOneMap site. Once datasets are superseded and turned into the electronic archiving, 3–5 people take care of the formal curation and preservation process.

Edward Curry: What are the uses of the curated data? What value does it give?

Joe Sewash: Curation of superseded data is based on an appraisal analysis translated to records retention schedule that is guided by the North Carolina General Statute. A significant deliverable of the GeoMAPP project was the development of a business planning toolkit; this toolkit encourages the practitioners to identify and accumulate data and information related to value of preserved datasets in support of return on investment calculations.

Use cases that we see tend to be around two big areas. The first is the recognition and acknowledgement of the records retention schedule that presumably is based on business analysis, not on technology analysis. You need to appraise and identify what you deem as preservation worthy. The second area is risk abatement. Imaging that your business process is evaluation of development permits and all of your source data is in a digital environment. In the future the sources of the data, the reason for the approval and the circumstances behind it may be lost. It is much more cost effective to do the preservation work on the front end as opposed to having to go back and unearth or recreate what you are looking at. Data accessibility and provenance are very important aspects from a technical point of view.

Edward Curry: What processes and technologies do you use for data curation? Do you have any comments on the performance or design of the technologies?

Joe Sewash: The geospatial data preservation process is a series of manual steps. A significant focus is placed on metadata maintenance from the data producer to NC OneMap hosting, and on to the NC DCR for curation.

The data preservation archiving is the division of cultural resources. We are housed in the IT organisation and the data is handed to the preservation side, which has discovery and access tools.

Edward Curry: What is your understanding of the term "Big Data"?

Joe Sewash: Thinking about this question for this interview has proven elusive. Identification of what the scope of Big Data is, or what lies beyond the scope of Big Data is practically irrelevant. As a practicing geographer, any dataset within the rubric of Big Data also has an explicit or implied spatial context.

Edward Curry: What influences do you think "Big Data" will have on future of data curation? What are the technological demands of curation in the "Big Data" context?

Joe Sewash: From a public sector context, the realities of the current and projected economic conditions will enforce linkages of curation capacity with expected value or returns. Within the US, there are not many examples of public-private partnerships to support advancements in this area.

At least in a geospatial context a public-private partnership is the equitable acknowledgement of data as an infrastructure consideration as opposed to commonly held perspective of data as a commodity.

If you are familiar with the ordnance survey and their economic model, it more closely resembles a public-private partnership than anything that you can see in the US. I understand that tremendous cost and really the financial model that goes behind the investment to create and maintain these very large datasets. It is a very much more participatory process that you can find the public private partnership. In the US, you are very ingrained in terms of tax payer funded resources and the availability of those resources, that there is some fundamental misalignment in terms of the economics versus the policy. I was at a conference earlier this year and the Canadian province of Alberta in the mid-90s actually went through a very severe economic downturn where the province was going to shut down their geospatial activities because they could not afford them anymore. And because Alberta is very resource extraction intensive, those industries came to the government and proposed a public-private partnership. The utility industry, the timber industry and the extraction industries came forward and said we need this data, we need a plan, we can support the financing because this is effectively a life blood in helping us doing our work and it has endured over the past 10–15 years.

The public sector has specific business needs. In the US system the states really perform a nexus between federal agencies and national datasets and local governments and very granular datasets that they need in their business processes. The public sector has certain business processes in the provision of public services that makes their need as germane as the private sectors, but also the integration piece and management piece. Regarding the economics and the pragmatic side of the technical implementation side, you could find ways to split of that cross of the partnership.

About the BIG Project

The BIG project aims to create a collaborative platform to address the challenges and discuss the opportunities offered by technologies that provide treatment of large volumes of data (Big Data) and its impact in the new economy. BIG's contributions will be crucial for both the industry and the scientific community, policy makers and the general public, since the management of large amounts of data play an increasingly important role in society and in the current economy.

CONTACT

Collaborative Project
in Information and Communication Technologies
2012 — 2014
General contact
info@big-project.eu

Data Curation Working Group
Dr. Edward Curry
Digital Enterprise Research Institute
National University of Ireland, Galway
Email: ed.curry@deri.org

Project Coordinator
Jose Maria Cavanillas de San Segundo
Research & Innovation Director
Atos Spain Corporation

Albarracín 25
28037 Madrid, Spain

Phone: +34 912148609
Fax: +34 917543252
Email: jose-maria.cavanillas@atosresearch.eu

Strategic Director

Prof. Wolfgang Wahlster
CEO and Scientific Director
Deutsches Forschungszentrum für Künstliche Intelligenz

66123 Saarbrücken, Germany

Phone: +49 681 85775 5252 or 5251
Fax: +49 681 85775 5383
Email: wahlster@dfki.de

Project co-funded by the European Commission within the 7th Framework Programme (Grant Agreement No. 318062)