**Access Innovations**

**An Interview with Margie Hlava**

**Editors note: To view the original source for this article see: http://www.arnoldit.com/search-wizards-speak/access-innovations.html**

|  |  |
| --- | --- |
| Margie Hlava of Access Innovations | In the last five years, controlled term indexing and rigorous taxonomy development moved from the specialized world of scientific, technical, and commercial database publishing to the mainstream. Overnight, self appointed experts advertised themselves as “experts” in taxonomy development. One prominent “expert” hired me to write about indexing methods applied to general business information. I later discovered that the newly minted “experts” were recycling content I created for a Web site article. No problem because I do work for hire, but the incident made it clear that “experts” in taxonomy related consulting were often more like a polyethylene shibbóleth.  I had an opportunity to meet with Margie Hlava, founder of [Access Innovations](http://www.accessinn.com) in early July. The full text of the interview appears below. |

**What brings you to Louisville, Kentucky?**

I continue to be active in the Special Libraries Association. I am in town for a meeting, and I wanted to make time to talk with you about taxonomies.

I’m flattered. Thanks. What’s the background for Access Innovations. You were founded in 1980, right?

Close. Access Innovations started in 1978. In 1973 I was a graduate student working for a NASA installation, called TAC.

**What’s that acronym stand for?**

Technology Applications Center (TAC). I moved from Information Engineer to Information Director with a team supporting searching of online databases which were quite new at that time. I was involved in a scientific and technical publication publications program which included Heat Pipe, Solar Energy, Wind, Remote Sensing and Hydrogen Energy Bibliographies.

**That sounds like work that is going on today.**

Yes, but energy research, like online information, reaches back decades. There is some wheel reinvention underway today.

So I was involved in work with the US National Energy Information Center. I became the Information Director at NEC and then got involved at the state level as well.

I thought there was much more that could be done in searching and building databases without the university and government bureaucracies. So I left my botany masters degree on the “Revegetation of Mine Tailings” behind and started Access Innovations with six forward-looking engineers and information scientists.

We became Access Innovations.

**Remarkable. What types of problems have you been resolving?**

The information problems have been popping up like the Whack a Mole game. [Our clients](http://www.accessinn.com/clients/index.html%20) continue need to know what they have, where to find it and be able to reuse, repackage that information on demand.With the explosion of “big data”, new solutions are needed for transforming, parsing, and accessing information. We started a services company to build databases and support publishing programs for others. We continue to do this because big data, social content, and rich content require fresh and innovative thinking. That’s what we do.

**Let’s go back in time. What was one of your first projects?**

Our first customer was [Control Data Corporation](http://en.wikipedia.org/wiki/Control_Data_Corporation) and we built Local Government Information (LoGIn) and Renewable Energy systems for them. That has evolved to organization, workflow analysis, specifications for normalizing markup, DTD [document type definitions] / schema creation, digitizing, and tagging.

**So you were active in structured document tagging from computer aided logistics and standard generalized markup language from their beginnings?**

Yes, today many “experts” are rediscovering fire. We have moved into laser fusion, but if you don’t understand the engineering behind taxonomies, much time and money are wasted floundering around.

You focus on smart software in your products and systems. Correct?

Partially correct. We do use next generation software methods. **But**, we have always valued the human intellectual effort highly and so constantly search for ways to make people more efficient and the tasks less clerical and more intellectual.

The combination of needing to handle other people’s data to the depth, breadth, accuracy as they do while insuring quality and consistency leads us to continuous automation and improvement processes.

We understand automated tagging. We understand the human value. We try to craft a solution that meets the specific requirements of the client without losing site of what each approach delivers.

**How does your firm fit into the landscape of search and retrieval?**

I have always been fascinated with logic and the application of it to the search algorithms was a perfect match for my intellectual interests.

When people have an information need, I believe there are three levels to the resources which will satisfy them.

First, the person may just need a fact checked. For this they can use encyclopedia, dictionary etc. Second, the person needs what I call “discovery.” There is no simple factual answer and one needs to be created or inferred. This often leads to a research project and it is certainly the beginning point for research. Third, the person needs updating, what has happened since I last gathered all the information available. Ninety five percent of search is either number one or number two.

These three levels are critical to answering properly the user questions and determining what kind of search will support their needs. Our focus is to change search to found.

**When did you become interested in text and content processing?**

While I was at TAC, I was curious about how the systems worked. I wanted to become a sponge, to learn as much as I could. I just immersed myself in the algorithms for the different search systems and what made them work.

In the late 1970s, it was dawn for the online industry and the principles were very willing to share information and even talk about how things could be done better. Meetings like the [Cranfield Institute](http://en.wikipedia.org/wiki/Cranfield_Institute) and [ASIDIC](http://www.asidic.org) brought together lead people to debate effectiveness of different approaches. We then put together teams to test the results immediately.

More than 30 years ago, the market was not yet swayed by a few government granting agencies with a particular notion on which systems they preferred. Academic researchers worked side by side with industry, not in competition with it. The field was open to different ideas, embracing different technical sectors, and committed to rapid testing. Those were fun, vibrant heady times.

**Can you describe how you developed your love/interest/expertise in content processing?**

I spent about 20 hours per week searching from 1973 to 1978. That was a lot of time on the teletype and then [Telnet](http://en.wikipedia.org/wiki/Telnet) and then long distance dial up modems. So we keyed and taped the searches in advance. We then sent what are called queries in batch bursts over the comms link to save connection time.

What we did now looks primitive, but our approach was pushing steadily to the reality we have today. Our approach also really pushed the systems. We needed to know and anticipate the logic and have the options available depending on the results from the search.

Online prints.

**What’s an online print?**

Oh, you have youthful readers, don’t you? An online print is a digital file that is output either on a screen or a line printer as a complete record or document. If the online print were sent to a thermal paper, the output would become unreadable quickly.

So prints of the records were sent to us by mail. I loved testing the commercial and government systems. I wanted to know how to derive the greatest value from our time and fees.

I loved figuring out the glitches as well. Today our curiosity might put me in the same category as a “hacker.” But online was new and we were trying to learn. The vendors like Dialog, BRS, Orbit Data Star and others definitely paid attention when I asked a question or was on a panel.

My personal Dialog command is "PRINT CANCEL" from when one of my staff put in the term "OXIDE" into Chemical Abstracts and then commanded the system to "PRINT ALL"

I knew the commercial vendors’ systems as well as their technical staff. Because I was looking at multiple systems, I had colleagues tell me that I knew multiple systems better than almost any other person fascinated by online information access.

Then when I went on my own, I continued to do searching for others. I also started building data files for ourselves and for commercial publishers. Today we are good at online and findability because we know how information retrieval works for users. Over one-third of the files on the ProQuest / Dialog service today have come through our shop.

**So you are not a newcomer to online and you won’t be recycling the information in my blog to learn about search?**

We read your blog. We just chuckle at the folks who think that saying the word “taxonomy” magically makes it possible for them to crack these quite difficult aspects of indexing. What looks easy is one of those facets of information technology that are among the most difficult challenges in computer science. How does one find a document relevant to a query when the words in the query do not appear in the document?

**Why is indexing becoming a hot topic?**

Indexing, which I define as the tagging of records with controlled vocabularies, is not new. Indexing has been around since before Cutter and Dewey. My hunch is that librarians in Ephesus put tags on scrolls thousands of years ago.

What is different is that it is now widely recognized that search is better with the addition of controlled vocabularies. The use of classification systems, subject headings, thesauri and authority files certainly has been around for a long time.

When we were just searching the abstract or a summary, the need was not as great because those content objects are often tightly written. The hard sciences went online first and STM [scientific, technical, medical] content is more likely to use the same terms worldwide for the same things.

The coming online of social sciences, business information, popular literature and especially full text has made search overwhelming, inaccurate, and frustrating. I know that you have reported that more than half the users of an enterprise search system are dissatisfied with that system. I hear complaints about people struggling with Bing and Google.

**How does indexing help solve this issue of user access and satisfaction?**

That’s the $64,000 question. Indexing enables accurate, consistent retrieval to the full depth and breadth of the collection. This does not mean that the statistics-based systems the government loves so much, will go away but, they are learning to embrace the addition of taxonomy terms as indexing.

To answer your question, relevant metadata, tagging, normalization of entity references and similar indexing functions just make it easier to allow a person to locate what’s needed. Easy to say and very hard to do.

**There has been a surge in interest in solving every information problem with a taxonomy; that is, indexing with terms other than key words in a document. On the surface, this seems to be gaining traction but on the other my research reveals that people are still quite unhappy with enterprise and Web search \* if \* those people are good searchers. Other people don't seem to know the difference, so why bother with a taxonomy?**

Search is like having to stand in a long line waiting to order a cold drink on a hot day. So there will always be dissatisfaction because “search” stands between you and what you want. You want the drink but hate the line.That said I think the reason controlled indexing (taxonomy or thesaurus) is so popular compared to the free ranging keywords is that they have control. They make moving through the line efficient. You know how long the wait is and what terms you need to achieve the result.

**Access Innovations has been investing in search, content processing, and text analysis for several years. Where does research deliver benefits to your clients?**

I mentioned that we are keenly interested in assisting the knowledge worker so that they can use their brains, not repetitive clerical tasks that can be automated. We are dedicated to that.

In order to make systems smarter and humans more efficient, we do a lot of investigation into the way the brain processes language and individual learning styles. The question of how we think of things and what mental paths we follow when we try to find something is important. How does the brain determine where to find something, whether it is stored in the house somewhere or in a computer. What will make finding the information more straightforward is the mirror of where and how we stored it in the first place.

Some people can find things easily in haphazard stacks piled on the floor. I am not one of those people. Other people prefer structured paths. We are enhancing thinking about the best ways to refine taxonomic strategies and the implementations of taxonomies. We want to take what I call regular or unstructured data and then add structure so people can find the item once it is stored regardless of the person’s mental “picture” for finding the needed information.

**What fields of inquiry do you consider relevant to this problem?**

Historically my areas of research are natural language processing, controlling language and how to enhance the search experience.

**Okay, to back up. What’s the payoff for your clients?**

The clients benefit by our knowledge of the bleeding edge so they can be cutting edge with the implementations and applications we provide to them.

We have recently been doing a lot of research on “Term Mining”. “Text Mining” has been around for a long time. Text Mining and involves processing large numbers of data sets to find the co-occurrences, generate relationship maps, and similar ways of grasping the data. Some of the methods are interesting but not very accurate.

What if we did conceptual extraction using the thesaurus terms and their synonyms using a rule base. Armed with these metadata, we could then look for occurrences. Would we be able to access with more precision? Would we have more detailed data maps? I think so.

So far we have found excellent results and have been able to use large data sets. We process several million records and then using our methods, we boil then down to trend analyses, data forecasts, and similar types of reports or outputs.

Our method often yields results our clients would not have believed without the data to back up the assertions or findings out approach generated.

**Many vendors argue that mash ups are and data fusion are "the way information retrieval will work going forward. What’s your take on this most recent “next big thing”?**

I like the mash ups, data fusion and visualization options. Each can lead to beautiful results making the interpretation of data easier. Visual approaches were not possible with what I call “flat file results.” Other newer technologies like linked data and personalization could be in this category as well.

That said, these new presentations of results depend even more heavily on well-formed data underneath. These methods use controlled vocabularies. Some of those vocabularies are conceptual, and others are geographic locations with latitude and longitude attached.

If the data underneath are poorly tagged, then the mashup will reflect nothing meaningful. Rotten data gives rotten results.

For Access Innovations, the need for well formed data imparts greater impetus for clients to get their backfiles data well tagged and maintained continuously via updates.

With good data, our clients are in a good position to do an incredible array of new and interesting things. Clients only have to invest once to see substantive results. Since our technology enables creating that structure the data for them at low cost, the mash up trend is really good for our company. Mash ups help take search to a new level.

This may be the reason indexing is such a hot topic now.

**Without divulging your firm's methods or clients, will you characterize a typical use case for your firm's search and retrieval capabilities?**

A typical client for us comes with a set of documents--structured , unstructured or both--and a need to structure them for storage and retrieval. Our clients normally want to have neutral format not tied to a particular software or hardware platform. Like most people who need to access information efficiently, the clients would like to find, reuse and often repackage parts of the information.

Our clients usually have three user communities: The people who create the documents, those who want to use them and those who need to manage them. Each of our clients’ needs are quite different. Some have the further complications of regulatory bodies or other legal requirements placed on the collections.

This is true of pharmaceutical firms but also of mining operations, hospitals, insurance companies and so forth. We meet these requirements. We structure the documents or wrappers for the content objects even if the content is not text. We use XML to ensure data portability. We use Java for platform independence. We use [Unicode](http://unicode.org/standard/WhatIsUnicode.html%20) for all data allowing special characters and multilingual applications. Then we structure the data with appropriate metadata to help ensure that the items of information can be found through various systems.

Our approach frees the client from any artificial constraints and allows the client to move quickly or deliberately as they choose to new technologies and applications.

One more thing. We keep very current with the activities in the standards worldwide and insure that we continue to meet and exceed those standards whether they are from the [ANSI](http://www.ansi.org/) / [NISO](http://www.niso.org/home/), [ISO](http://www.iso.org/iso/home.html), [W3C](http://www.w3.org/). [IFLA](http://www.ifla.org/), government agencies, and other sources.

**What are the benefits to a content generator when working with your firm?**

Besides the fun of using their minds instead of being mired in clerical tasks? [Smile]

As I said earlier, the data processed by our systems are flexible and free to move. The data are portable. The format is flexible. The interfaces are tailored to the content via the DTD for the client’s data.

We do not need to do special programming. Our clients can use our system and perform virtually all of the metadata tasks themselves through our systems’ administrative module. The user interface is intuitive. Of course, we would do the work for a client as well.

We developed the software for our own needs and that includes needing to be up running and in production on a new project very quickly. Access Innovations does not get paid for down time. So our staff are are trained. The application can be set up, fine tuned, deployed in production mode in two weeks or less. Some installations can take a bit longer. But as soon as we have a DTD, we can have the XML application up in two hours. We can create a taxonomy really quickly as well. So the benefits, are fast, flexible, accurate, high quality, and fun!

By the way, having come out of the NASA world we are interested in technology and [technology transfer](http://www.nasa.gov/open/plan/technology-transfer.html). The colors we chose for the display screens are not flashy but they are the colors chosen by the US Air Force as those easiest on the eyes of people looking at screens for a long time.

**How does an information retrieval engagement with your firm move through its life cycle?**

Great question. Access Innovations is content focused. The technology is there for the content, not the other way around. For our life cycle, there are a number of pragmatic steps. First, we need to have the data well formed (as the XML people say). We want the data well tagged and in consistent format. Once that is done the rest is pretty easy.

Most of the problems people have in implementations are due to bad data. It is GIGO, garbage in, garbage out problem. If we cannot ensure that the data are clean and appropriately tagged, we will not accept the project.

The flip side is that with that clean, well formed, semantically enhanced with metadata tags from the taxonomy data, our client has many more options. For example, content can be repurposed or new platforms supported quickly, easily and economically. Most clients stick with us because of the flexibility, quality and accuracy of our work.

So the life cycle, has these steps:

1. [Create DTD](http://www.w3schools.com/dtd/default.asp) – or use clients and fix any parsing errors
2. [Clean data and load to XIS](http://www.dataharmony.com/products/xis.html%20) (XML intranet System for content creation)
3. Create taxonomy (using [Data Harmony Thesaurus Master](http://www.dataharmony.com/products/thesaurus_master.html))
4. Apply taxonomy and other metadata to data set automatically (Using [Data Harmony MAI](http://www.dataharmony.com/products/mai.html))
5. Export file as [XML](http://en.wikipedia.org/wiki/XML)
6. Load to search system; for example, [Lucene](http://lucene.apache.org/java/docs/index.html), Perfect Search, or any other system
7. Tailor the interface.

**Is this UX or user experience?**

Yes, the jargon is now [UX](http://en.wikipedia.org/wiki/UX). So we work on the search interface for auto completion using taxonomy terms and synonyms. We work to serve contextually appropriate related terms to expand search, serve contextually appropriate narrower terms to refine search, display “more like these” using similar index sets for documents, and show the hierarchy with term counts as a way to navigate the content, among other UX functions.

We then create visual search and [mashups](http://en.wikipedia.org/wiki/Mashup_(web_application_hybrid) using the record metadata. The last step is that when new data are added, we enhance those data with metadata, insure it is well formed and repeat from the fifth step.

**One challenge to those involved with squeezing useful elements from large volumes of content is the volume of content AND the rate of change in existing content objects. What does your firm provide to customers to help them deal with the volume problem?**

One of the wonderful things about our data Harmony MAI system is that it handles large volumes of data very well. It is as scalable as the Internet because it uses that base technology.

**What does MAI stand for?**

Oh, machine assisted indexing. Sounds old fashioned now so we just use the initials.

We can handle “big data” and changing data. We designed MAI to handle large volumes of digital content. In fact that is part of the reason for it. We needed something to help us with the editorial process of determining what a content object is about in a “big data” context.

Concepts covered well today will always change over time. Therefore, we needed to deal with changes to a knowledge domain. When a domain remains static, it is dead. Therefore, we must build the ability to handle change into the system and workflow. We have automated most of the processes so that volume is not the problem. The hard problem is finding. Searching has many methods. Finding is very hard.

**How do you cope with language drift?**

Changes to the terminology used in a field of knowledge are also constant. As you asked, “language drift.”

Once the records are tagged with the current terms, the records are stored in a data base. Most database system administrators are not enthusiastic when the the data set has to be rebuilt due to term changes.

Although we have found some partners where term change is not a problem, we have many clients who have to keep pace with change. So we concentrate on the user interface side. Remember I noted the interface work that can include synonyms, and other kinds of terms.

If you cannot rebuild the search index, you can build the newer terms into the interface while keeping the older terms used as well. Our approach provides additional flexibility. We keep track of those changes in the Thesaurus Master. Then whether the user is using [Microsoft SharePoint](http://sharepoint.microsoft.com/en-us/Pages/default.aspx), Lucene, [Autonomy](http://www.autonomy.com), [Endeca](http://www.endeca.com), or other retrieval systems, the user of one of these systems can still find the current terms. We harvest those terms in a number of ways to insure currency.

**Another challenge, particularly in professional intelligence operations, is moving data from point A to point B; that is, information enters a system but it must be made available to an individual who needs that information or at least must know about the information. What does your firm offer licensees to address this issue of content "push", report generation, and personalization within a work flow?**

In our case, each system begins with the thesaurus (taxonomy). We then use MAI to tag the data as it feeds in to the system. Individuals who break the law or public relations professionals know exactly how the co-occurrence algorithms work. Most are simple to circumvent. Tagged content is much more difficult to trick.

Recently there has been much press about how Google tailors results, and it seems to be a big surprise to many people. But really that is what they system was built to do and why it is so successful.

If instead we were to consistently tag data and have algorithms to keep track of the terms changes in use and morphology, we would be much more successful in tracking the data feeds of all types to a neutral terminology.

**Give me an example.**

Okay, for an intelligence professional we offer feeds based on profiles using the terms in the thesaurus in combinations appropriate to their cases. If they are tracking an organization or a person for example we need to track that information in the source language if possible to avoid the errors and many different ways to transliterate from one language to another. That is where the Unicode options used in our technology become particularly useful. We can have the source come in in one language and character set and be offered to the analyst in another language.

**How does your automated function respond/learn/adapt to new concepts for which there is no antecedent term?**

This to me is novelty detection. Finding new terms in use by the community is very common. People constantly invent new terms, call old things by new names, shorten the names, pronounce initialisms, etc. We can deplore it like the Castilian Institute in Spain or the XXX in France but people still do it. All this texting technology only makes it worse.

Novelty detection itself, although we have it, is expensive and imprecise. We deploy it with caution. Surprise thought: the world really isn’t that novel. Take science, for example. Most new discoveries don’t require a new term. They require a new way of looking at old things. So it is the conceptual environment for the terms that change.

We discover these terms two main ways in practice.

One is looking at the search logs to see what people are asking for, the other is looking at the incoming data and harvesting new terms not already in one of our dictionaries. Not top difficult.

What is more difficult is looking at the term combinations. It is looking at things in a new way. That would lead to new edges of science, hence new combinations. This is where the Term Mining applications come into play. This is a very interesting area!

**I am on the fence about the merging of retrieval within other applications. What's your take on the "new" method which some people describe as "search enabled applications", personalization, and filter bubbles? Do users even have to search anymore?**

You could consider this searching in new ways. Yes, people are still searching. They are just using a different interface to run the query. The underpinnings still involve search. More than ever, these new interfaces require that the underlying data is well formed and well tagged.

Put these new interfaces on top of poorly done data and the results can be really poor. The scary part of these systems are people are one more step removed and may not know how poorly they are being served by the system.

**There seems to be a popular perception that the world will be doing computing via iPad devices and mobile phones. My concern is that serious computing infrastructures are needed and that users are "cut off" from access to more robust systems? How does your firm see the computing world over the next 12 to 18 months?**

Like everyone else we have moved to supporting people remotely. There is the cloud (through a browser) or an SaaS approach (using a client on the user machine and main software on the host).

I believe is both suggest a general direction for the future. It does not mean that there is less computer power needed. In fact, more computing capability is needed as well as more band width. Since the advent of the iPhone, there are many more dropped calls now that there are more apps and pictures being moved on the internet. The system is struggling to cope with increased data flows.

This cloud and SaaS approach may mean that the users are freed from debilitating waits for service and project implementation using internal information technology departments. No one can usurp the marketing department’s servers for a higher priority jobs. New approaches mean enabling the user, not controlling access to data and systems.

We are seeing a very fast change from internal hosting to Web hosting of data and applications. I think that trend will continue. We embrace it. We offer monthly access to the applications using our servers. People use either cloud or the more secure SaaS options quite happily.

The cloud shift also means we do not have to deal with as many strange internal configurations, firewalls, and other challenges in getting a customer up and running. It also means the device they use on the client side can be flexible and small. We can have them online with their data and applications in hours.

IBM sells more services than hardware, BUT they still sell billions off dollars worth of “big iron” every year. More will be done in the cloud, but the shift won’t happen overnight so there will be customers who need on premises installations. We continue to offer cloud and on premises options to our clients. Regardless of the approach the client selects, we help make the customers content independent of the applications and platform, thus portable.

**Put on your wizard hat. What are the three most significant technologies that you see affecting your search business?**

I think I have mentioned these. First, there is the cloud shift. The move to the “Cloud” and SaaS will add greater potential for data sharing and service enhancement. I also believe that there will be more options for maintaining security and intellectual property rights, an interesting mix.

Second, I am enthused about Term Mining and associated ways to use enriched data for business intelligence, trends analysis and forecasting based on the internal data and data feeds associated with the business of an enterprise.

Third, Like others, I enthused about the appetite for taxonomies and other metadata enrichment. I think we have just scratched the surface on this one. It will form the basis of the mashups but also author networks, project collaborations, deeper and better retrieval.

**Where does a reader get more information about Access Innovations?**

I suggest that you visit out Web site at [www.accessinn.com](http://www.accessinn.com).

**ArnoldIT Comment**

For more than 25 years, Access Innovations has been provide ANSI-standard taxonomy and controlled vocabulary products and services. The firm’s taxonomy management system is widely used across professional publishing, commercial database publishing, and in dozens of the world’s leading research institutions. The company develops work flow systems tailored to the specific production requirements of its clients, provides consulting and production services to organizations worldwide, and maintains the line of Access Innovation software systems.

In the last few years, taxonomy has been turned into a buzzword. There are “experts” who have never built an ANSI standard compliant term list lecturing those interested in taxonomies. In my experience, the consumerization of the concept of “tagging” has created the findability failures that are prevalent in commercial, governmental, and non profit organizations today.

Avoid this failure from incomplete, skewed, or inaccurate information about controlled vocabulary lists, taxonomies, and ontologies. Access Innovation is one of the preeminent firms in this discipline. My recommendation? Talk with Access Innovations. When I worked in the commercial database business, the excellence of our indexing was, in part, a direct result of the contributions Access Innovations made to my units. Highly recommended.

*Stephen E Arnold, July 19, 2011*