NICS Symposium - November 2, 2005
"NICS and Metadata: Joys, Sorrows and Payoffs"

Session A: Creating Metadata for Your Data Sets, led by David Stevens and Claudia Coulton

Discussion Points

Discussion of crosswalks noted that while customers have an incentive to receive what they want (customized) data providers do not have an incentive to deliver customized formats. [Stevens did not mention, but feels strongly, that crosswalks by definition 'blur' details in the original file taxonomies that are presumed to have value that is lost through the crosswalk exercise.]

Discussion of the 'data web' topic focused on multiple levels of data quality; and it was suggested that Federal statistical quality standards and practices should be emulated at the state and local levels. However, I noted that Alan Tupek had said in earlier remarks that Census does not practice what it preaches.

Cavan commented on the "I want to see your data, but I do not want to share my data with you" challenge, and noted the practical problem of data overload dangers—too much data overwhelming potential customers who are not prepared to spend adequate time sorting through to find what they want/need.

Andy pointed out that the preliminary design of NICS would impose a reciprocity criterion for participation—that access to others' data is contingent upon cooperation in providing your data. A basic goal is to enable better benchmarking that can improve performance-driven management strategies.

David mentioned that the NICS use-cases are intended to add value by example, similar to what Jacqueline meant when talking about the best-practice county-level statistical systems.

Andy pointed out that NICS is envisioned as a market-making or market promotion intermediary—advancing value-added that may not even be perceived today through judicious development of reliable metadata.

Someone mentioned that commercial 'stove-pipe' software is not aligned with Federal statistical system needs.

Someone mentioned that business process data sources are transformed into statistical uses; and that these data 'sit' close to the business processes.

Suggestions

Jacqueline (NACO) suggested posting of “high performer” templates that might be used by others for comparison applications. She also noted that local control over data content is limited, and referred to the ‘silos’ challenge.

Someone asked whether there is a market-making opportunity here, which would allow data ‘owners’ to charge for access to their data.

Cavan commented on the building-block capacity that might be provided through NICS, offering a rapid-response to such events as Katrina.

Summary

David: we need to develop the idea of NICS as setting an agenda, lowering transaction costs, demonstrating value-added through use-cases, setting voluntary rules for adoption of meta-data approaches, incentives to improve customer understanding aligned with data provider interests in sharing, and a pivotal role in promoting convergence toward a voluntary consensus on meta-data standards.

The biggest challenge appears to be the **reciprocity** criterion or rule—it is not clear to me how NICS as an intermediary broker or market-maker can balance the desire to introduce many customers to many providers with the constraining rule that requires sharing of something to gain access to something else.

Some are ready, willing, and able to do metadata, but they need a system of rules and standards to follow:

- *Incentives to provide metadata* -- What is the business model for the data creator? What is their market?
- *Technology* -- Educate data creators about what is available and ask them what they need. Technology will follow if the tech people know what is needed. Technology is constantly changing -- NICS can reduce transaction costs by letting people know best practices and what is available.
- *Convergence* - NICS facilitation will bring about convergence. For example, everyone will start using the same terms. Volunteerism.
- *Use Cases* -- Demonstrate value added and reduce transaction costs. Educate about best practices
- *Reciprocity* -- what will it take to get me to put my data out there? Determine why metadata *isn't* out there.
- *Set the agenda* -- NICS can be the market maker

Session B: Building Metadata from Existing Data, led by Roderick Harrison and Shelia Denn

Discussion Points

Need to get at a standard, framework

With data that is highly idiosyncratic, only a few people can ever be an expert at that data. With data built/collected in a broader framework, more people can be experts at a datasets

What is the quality of this data & what faith should I have in it may be addressed by a system, a standardized process. Knowing that it went through a certain process.

We are never going to have all the data you're looking at, explained, so we have to be able to deal with accepting this fact

A chicken and egg scenario: we aren't ever going to get to X level until someone creates some standard, and people start to move towards it

Types of data: survey, administrative data, and then self-reporting/community data

We're aiming at two different purposes:
1. Give the expert user an expert answer, then these comprehensive standards are the types of standards we need
2. Novice, not statistically advanced user might just need a "NICS stamp of approval", or knowing that the data went through a standard process, and that process guarantees the dataset in question of it's

Novice versus expert users: how each arrives at an answer. We've seen (Patricia...from VA) Novice users typically enter the site via Google (pulled in by the metadata on the pages). Users find answers without going through a step-by-step process. Systems aren't necessarily self-contained anymore. So, we need to consider this when thinking about how novice users may find the answers we are trying to provide.

different users can stand different "levels of risk" - quick answers, ones that will guide big monetary/time investments --so the user has to determine what level of risk/quality/trust they are comfortable with or what minimum level they'll allow for the decision they are making

Should NICS should get into the quality certification business, because it is very expensive; if we were to do it, we shouldn't claim the data is trustworthy or risky -- we need to be precise about the multidimensional nature of why that data is

better or worse

Metadata isn't a document that says whether to trust a data source or not trust a data source, it's highly technical. The onus of are we going to trust this -- is on the analyst.

In 15 years, we'll have the technical answers about how to combine the data with the metadata by better means. The right now, addressing needs of users now - is some clearing document, the ISO 9000; the Taeuber/Smith - clearing house

Technology and culture is changing how we access information - people Google instead of going to their library to get answers.

Suggestions

Can we set up a threshold to distinguish what metadata is needed for what data? With self-reported data (like much of EPA data) - it lacks a frame, has little metadata

ACCRA has a standard that they use to include or exclude data

FedStats has a question mark link with all tables that opens up the metadata on the numbers being generated

Use multidimensional index to make a threshold that makes data NICS-certified/ready/ok

Suggestion to proceed with a "Cindy Taeuber" type list (see fig 2. p 29 on NICS and Metadata white paper).

Tupek handout, Taeuber/Smith paper is almost a chronological process that could fill this "standardized process"

Even with googling answers, users could find an answer via Google, and then NICS could serve as a means to compare/evaluate how good that answer is, or pull up other related answers e.g. vacancy rates--Google gets a vacancy rate for VA, but NICS can pull up all the different measures for vacancy and the user or the system would rank how "trustworthy" that answer is.

(A.Lomax) I want to make people *more* uncomfortable with using data. They are already *too* comfortable with using data.

(Deberry) That's why I like the idea of a simplified questionnaire. Some metadata with a questionnaire, that's at least a start to get at the concept of quality.

(Lomax) My sense is that folks who get numbers from the federal community

think that it's good.

Would our analysis of metadata to give a "gold star" standard be controversial? Would some of the metadata be open to interpretation, discussion?

I think you can create a set a guidelines, with fuzzy boundaries, that sets a minimum standard

Just focusing on data intermediaries will be a more cost-efficient way that educating everyone.

Reinvent the statistical publishing paradigm: XML, HTML formats for tables prevents users from having to screen scrape; interoperable; "living" by making it open to add

BASIC METADATA: which elements?

- Too many items in the "double-starred" items in Taeuber/Smith. For the novice data users, what is the basic, should we set our lowest common denominator
- the FGDC standard is nothing compared to the Taeuber/Smith paper, so don't worry too much about what you have here
- Double-star the geographic levels included in the survey, sample, data.

Summary: Action Points

- We should proceed with some version of the Taeuber/Smith, Tupek list (from federal register, included in Taeuber/Smith white paper) to create a "data quality checklist"
- NICS should use some version of that to adapt at the federal level
- Regarding Core/Critical meta elements: distinguish by level of user, and type of data
- Data Intermediaries: biggest metadata challenge is for local data, the federal information is largely standardized.
- Harmonize the naming conventions, definitions (e.g. def of poverty)

Biggest Challenges:

- at local: biggest challenges are administrative data
- self-reported w. selection bias surveys are the biggest challenges
- Approach could be to utilize the existing federal data communication, development and standards building procedures, that NICS could plug into