2012 Federal CASIC Workshops

CASIC

Computer Assisted Survey Information Collection

Survey Uses of Metadata

Metadata: To Boldly Go Where No One Has Gone Before?

Pascal Heus
pascal.heus@metadatatechnology.com

Vice-President
Metadata Technology North America
http://www.mtna.us

Executive Manager
Open Data Foundation
http://www.odaf.org

www.openmetadata.org

metadata technology NORTH AMERICA

Open Data Foundation

CASIC

2012 Federal CASIC Workshops

Computer Assisted Survey Information Collection

Survey Uses of Metadata

Metadata: To Boldly Go Where No One Has Gone Before? PART1: Captain's Log

Pascal Heus
pascal.heus@metadatatechnology.com

Vice-President
Metadata Technology North America
http://www.mtna.us

Executive Manager
Open Data Foundation
http://www.odaf.org

Robert M. Groves Blog – 02/2012

- Director of US Census Bureau
- February 2012: Blog entry on "National Statistical Offices: Independent, Identical, Simultaneous Actions Thousands of Miles Apart"
  - http://directorsblog.blogs.census.gov/2012/02/02/national-statistical-offices-independent-identical-simultaneous-actions-thousands-of-miles-apart/
- Recent summit by leaders of statistical agencies from Australia, Canada, New Zealand, United Kingdom, and the United States
- To identify common challenges and share information about current initiatives
- Initiative lead by Brian Pink from Australian Bureau of Statistics
- Perceive the same likely future challenges and making similar changes

Summit - Vision ingredients

- Digital data increases faster outside NSO than within
- Traditional surveys/censuses becoming less attractive
- Blending data from multiple admin and other sources with surveys/censuses must become major focus
- Requires: Efficient record linkages, master universe frames, modern statistical modeling
- Agencies need to develop the analytical and communication capabilities
- Growing demands from researchers and policy-related organizations to analyze the micro-data

Summit - Implications for agencies

- Traditional functional separations among surveys/censuses are not well-fitted. Need to unify collection processes.
- Need generalized IT systems
- Program agencies need support from statistical agencies staff
- Need to high-speed big data
- Efficient and sophisticated imputations procedures are needed to combine data
- Use of statistical modeling for statistical estimation, to provide more timely and small area estimates
- The agencies are inventing new ways to give secure access to micro-data for legitimate research purposes

International Log

- International Household Survey Network
  - NSO in developing countries (100+ countries) (DDI-C)
- Australian Bureau of Statistics
  - REEM, IMTP, BigData (DDI/SDMX driven institutional data management framework)
- Canada Research Data Centre Network
  - Secure access to Statistics Canada datasets (DDI driven)
- Data without Boundaries (DwB - EU)
  - 28 partners, 20 countries (research infrastructure) (DDI/SDMX)
- …

US Log

- Statistical Community of Practice and Engagement (SCOPE)
  - Shared environment / infrastructure for US statistical agencies
  - Data protection / disclosure control
- Cornell NSF-Census Research Node: Integrated Research, Support, Training, and Data Documentation
  - Facilitates access to detailed metadata on (1) restricted–access data from outside and (2) RDC Public-use datasets inside restricted-access areas
  - Metadata to support disclosure control / review
  - Foster capture of user-generated metadata
- NORC Data Enclave
  - Secure virtual remote access to sensitive data
  - http://www.dataenclave.org

Standards Log

- General Statistical Business Process Model (GSBPM) 4.0
- Generic Statistical Information Model (GSIM)
  - information model to complement the GSBPM
  - Version 0.3 available for comments
- Data Documentation Initiative (DDI)
  - DDI-Codebook 2.5 released Jan 2012
  - DDI-LifeCycle 3.2 planned for 2012
- Statistical Data and Metadata Exchange (SDMX)
  - Version 2.1 released (ISO ongoing)
- RDF
  - Serialization of SDMX available, DDI in progress

Data.gov Log

- Open data / data.gov
  - Began in 2009 in US
  - Spread to other countries
    - United Kingdom, Canada, Australia, New Zealand,…
  - and International organizations
    - World Bank,…
- Driven by
  - Demand for data
  - Transparency
  - Politics / Semantic web / Linked Data
- But presents new challenges to statistical agencies who are asked to deliver data while at the same time respecting statistical principles, in particular around privacy and disclosure control

IT Industry Log

- XML / SOA (service oriented architecture) continues to be the industry successful and standard model
- Big data buzz
  - Big Data warehouse (structured data)
    - Share nothing multi-node column oriented databases
  - NOSQL / Hadoop (unstructured data)
    - But don't use unless truly needed
  - + do you really have "big" data (Tb-Pb)?
- Linked Data / RDF
  - Coined by Tim Berners Lee around Semantic Web (but concept has been around for along time)
  - But this requires linkable data….
  - RDF: SDMX → RDF, DDI → RDF, Graphs, SKOS

Log Analysis

- Lots of great ideas and vision out there
  - But we lack a roadmap
- Lots of thing are happening, but what's reality vs fiction?
- Metadata and related technologies are fundamental components
- What is actually possible?
  - And why it's not happening…..
- What's step one? How many steps do we need to take?
- Short term vs long term?
- What's the big picture?

CASIC

2012 Federal CASIC Workshops

Computer Assisted Survey Information Collection

Survey Uses of Metadata

Metadata: To Boldly Go Where No One Has Gone Before? PART2: What do we know?

Pascal Heus
pascal.heus@metadatatechnology.com

Vice-President
Metadata Technology North America
http://www.mtna.us

Executive Manager
Open Data Foundation
http://www.odaf.org

What do we know?

- What is metadata
- What metadata is good for
  - which is much more than what most know about
- IT technologies are widely available (XML, SOA, etc.)
- Metadata and related technologies are successfully being applied in many domains (i.e Internet)
- Standards for socio-economic and health data are available
- Tools are available (or rapidly emerging)
- "Metadata" is widely recognized as the right thing to do, but what does it mean?

What is Metadata?

It's not just 'data about data' …

COURTESY: GENE ILLS

Nutrition Facts
Valeur nutritive

Per 1 bowl (300 g) / Pour 1 bol (300 g)

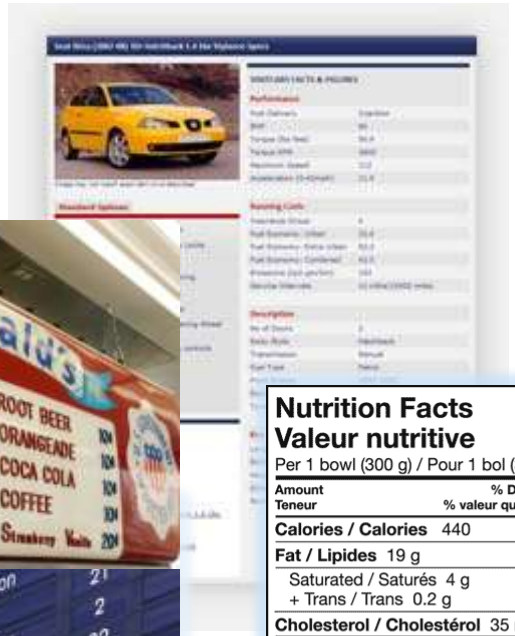
Amount / Teneur	% Daily Value / % valeur quotidienne
Calories / Calories 440	
Fat / Lipides 19 g	**29** %
Saturated / Saturés 4 g + Trans / Trans 0.2 g	**21** %
Cholesterol / Cholestérol 35 mg	
Sodium / Sodium 860 mg	**36** %
Carbohydrate / Glucides 53 g	**18** %
Fibre / Fibres 4 g	**16** %
Sugars / Sucres 6 g	
Protein / Protéines 15 g	
Vitamin A / Vitamine A	45 %
Vitamin C / Vitamine C	4 %
Calcium / Calcium	20 %
Iron / Fer	20 %

… we need metadata to understand what things are about…

http://www.metadatatechnology.com
@metadatatechnology.com

Everyday Metadata

"Human Readable" Metadata



Nutrition Facts
Valeur nutritive

Per 1 bowl (300 g) / Pour 1 bol (300 g)

Amount / Teneur	% Daily Value / % valeur quotidienne
Calories / Calories 440	
Fat / Lipides 19 g	**29** %
Saturated / Saturés 4 g + Trans / Trans 0.2 g	**21** %
Cholesterol / Cholestérol 35 mg	
Sodium / Sodium 860 mg	**36** %
Carbohydrate / Glucides 53 g	**18** %
Fibre / Fibres 4 g	**16** %
Sugars / Sucres 6 g	
Protein / Protéines 15 g	
Vitamin A / Vitamine A	45 %
Vitamin C / Vitamine C	4 %
Calcium / Calcium	20 %
Iron / Fer	20 %

"Machine-actionable" Metadata

Metadata has been successful

YAHOO! msn AOL iGoogle

Weather Forecast

Compact | Classic | F

1°C
Mostly Cloudy

Location	Today	Tomorrow	Tuesda
Washington, DC Mostly Cloudy	5° / 0°	10° / 4°	7° / 0°

Top Stories from AP

- Obama aide promotes job plan, warns automakers - 1 hour ago
- Sources: Gov't working on Citigroup rescue plan - 37 minutes ago
- Police: Wife shot and killed at New Jersey church - 40 minutes ago
- Official: Richardson to be commerce secretary - 4 hours ago
- AP IMPACT: Govt pays millions for unapproved drugs - 55 minutes ago
- Astronauts tinker with urine-to-water machine - 2 hours a
- Wall Street braces for another pivotal week - 1 hour ago
- 'Twilight' takes $70.6M bite out of box office - 3 hours ago
- Manning shines, Giants hold off Cardinals 37-29 - 52 minutes ago
- TO snags 7 catches as Cowboys defeat 49ers 35-22 - 56 minutes ago

Scoreboard

Yesterday | Today

NFL — All games

Sunday, November 23, 2008

Washington	20	Final
Seattle	17	

AAPL	82.58	+2.09	+2.60%
AFFX	2.16	-0.22	-9.24%
ARIA	1.27	-0.11	-7.97%
AMZN	37.87	+2.84	+8.11%
EBAY	12.01	+0.84	+7.52%
GOOG	262.43	+2.87	+1.11%

Currency Converter

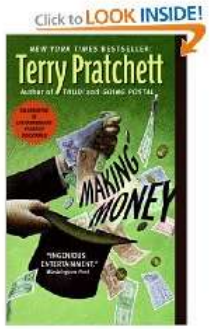
Convert 1 U.S. Dollar (USD)
into Japanese Yen (JPY) Convert

Currency	U.S. $	¥en	Euro	U.K. £
1 U.S. $ =	1	95.05	0.7936	0
1 ¥en =	0.0105	1	0.0083	0
1 Euro =	1.2601	119.7725	1	0
1 U.K. £ =	1.4891	141.5389	1.1817	

Quotes from Yahoo

amazon.com

Click to LOOK INSIDE!



Making Money (Discworld Novels) (Mass Market Paperback)

by Terry Pratchett (Author)
★★★★☆ (101 customer reviews)

List Price: $7.99
Price: $7.99 & eligible for **FREE Super Saver Shipping** on orders over $25.
Details
Special Offers Available

In Stock.
Ships from and sold by **Amazon.com**. Gift-wrap available.

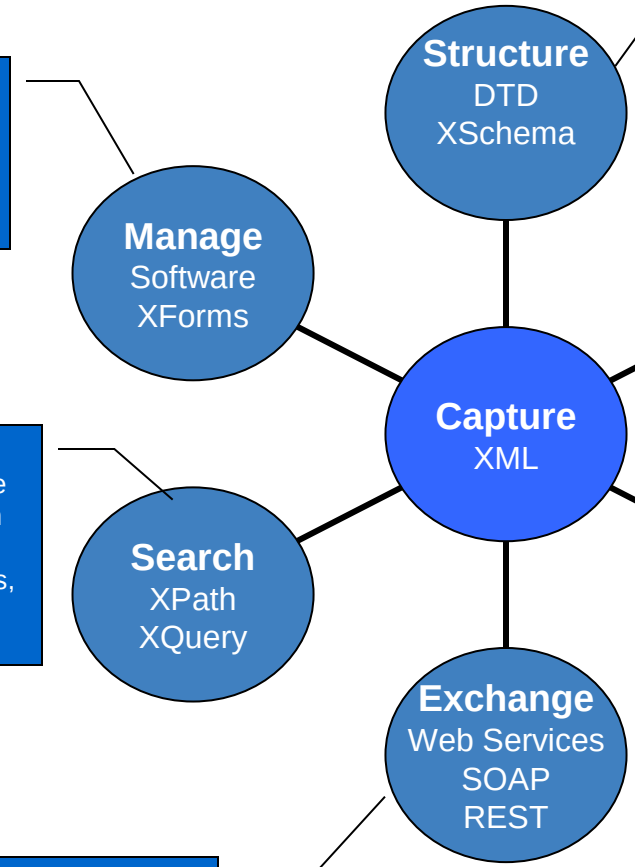
Want it delivered Tuesday, November 25? Order it in the next 21 hours and 9 minutes, and choose **One-Day Shipping** at checkout. See details

37 new from $3.48 10 used from $3.49

Why do we use or need metadata?

- Contextualize
- Discover/Search
- Promote/Advocate
- Document/Visualize
- **Automate, automate, automate!**
- Exchange (standards!)
- Secure/Protect
- To make sound decision, share knowledge, search

Technologies: XML (RDF)



Document **T**ype **D**efinition (DTD) and **XSchema** are use to **validate an XML document** by defining namespaces, elements, rules

Specialized software and database systems can be used to **create** and **edit** XML documents. In the future the XForm standard will be used

XML separates the metadata storage from its presentation. XML documents can be **transformed** into something else, like HTML, PDF, XML, other) through the use of the eXtensible Stylesheet Language, XSL Transformations (XSLT) and XSL Formatting Objects (XSL-FO)

Very much like a database system, XML documents can be **searched** and **queried** through the use of XPath oe XQuery. There is no need to create tables, indexes or define relationships

XML metadata or data can be **published** in "smart" catalogs often referred to as **registries** than can be used for discovery of information.

XML Documents can be sent like regular files but are typically **exchanged** between applications through Web Services using the SOAP and other protocols

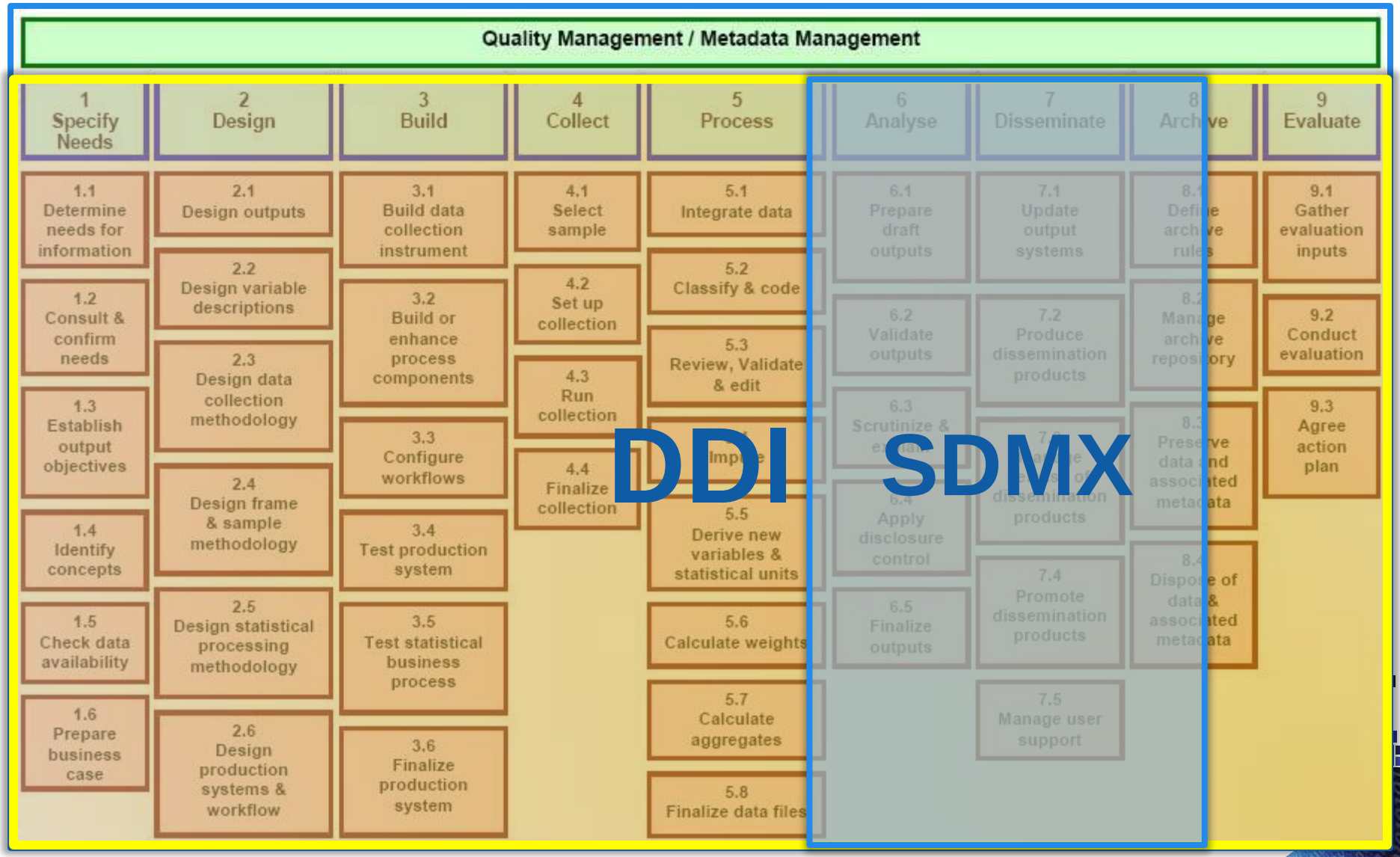
Data Technologies

- Structured (with various level of metadata)
  - Statistical / math packages
  - Relational Databases / Data Warehouse
  - Big data engines (Vertica, Greenplum, Teradata, Exabyte, InfoBright, …)
  - Spreadsheets (Excel, etc.)
  - ASCII
- Semi-structured / unstructured
  - NOSQL / Hadoop
  - RDF / Graph Databases
- File resources
  - File systems (not too good)
  - Controlled environments: iRODS

We have standards / frameworks

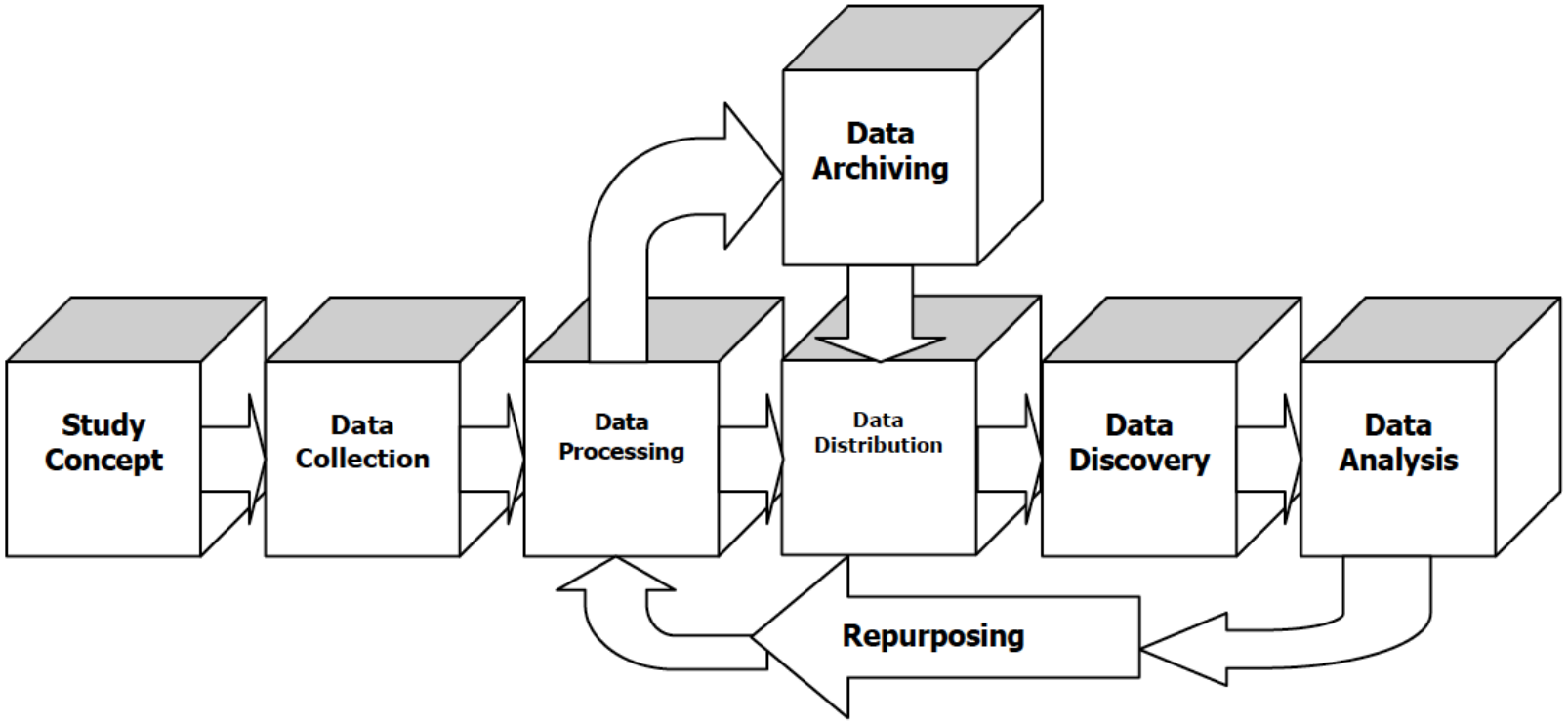
- METIS
  - Common Metadata Framework, General Statistical Business Process Management (GSBPM). Generic Statistical Information Model (GSIM), DDI/SDMX integration
  - http://www1.unece.org/stat/platform/display/metis/METIS-wiki
- Data Documentation Initiative (DDI)
  - Microdata or observation level administrative data
  - Maintenance: DDI Alliance
  - Expert statistician, researchers, data producer.
  - Two flavors: DDI Codebook (1.x-2.x) and DDI LifeCycle (3.x)
- Statistical Data Metadata Exchange Standards (SDMX)
  - Aggregated/Time series data
  - Maintenance: SDMX sponsors
  - Decision maker, casual user, economist, public, press
- ISO11179, Dublin Core, ISO 19115, etc.

DDI/SDMX and GSBPM



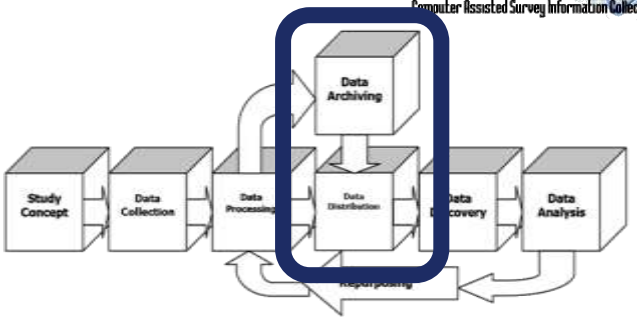
http://www1.unece.org/stat/platform/display/metis/Generic+Statistical+Business+Process+Model

DDI Lifecycle



DDI for Archive/Preservation

- Captures comprehensive information about surveys and their data



- DDI is widely used by national statistical agencies, data archives, research centers around the globe
- ASCII + DDI is also a powerful combination for long term preservation (non-proprietary text format)

From ASCII+DDI to other formats

A standard transformation can be applied to DDI XML to generate ASCII import scripts for various statistical, mathematical, office or database packages. Metadata can be inserted as comments.

Native software executes the script to import the data and create a file in the proprietary format.

ASCII DATA

script.sas → SAS → data.sas7bdat

script.sps → SPSS → data.sav

script.do → STATA → data.dat

script.r → R → data.r

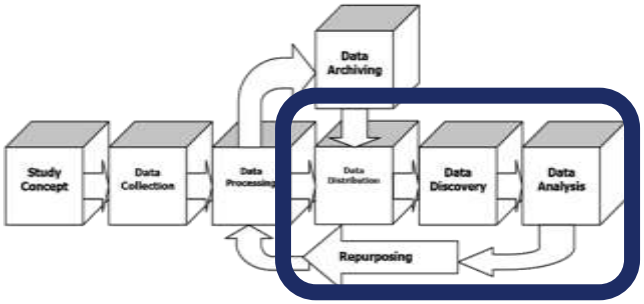
script.sql → Microsoft SQL Server, IBM, MySQL, ORACLE → Database

CSV → data.xls

…

Note that a single transform works generically for all ASCII+DDI datasets. Adding new packages only requires the development of a new transformation.

DDI for Discovery/Access/Analysis

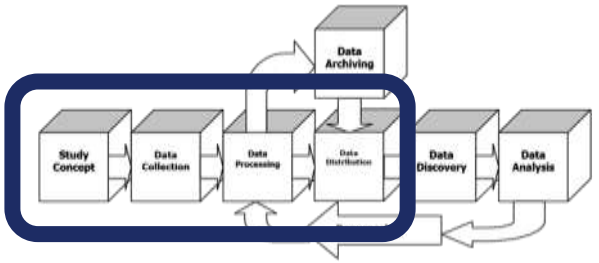


- Facilitates discovery through web services, portals, registries, subscription/notification, etc.
- Enable implementation of complex search engine and metadata mining tools
- Provide comprehensive information for users
- Can automate imports, transformations, custom documentation
- After the fact comparability
- Repurposing (adds new knowledge to the survey)
- Supports harmonization / data linkages

DDI for Production

- Designed for use from day 1 of a study or program
- Manage common metadata elements - such as concepts, universes, geography – across surveys or waves, or even agencies
- Supports classifications, question, variable, concept banks
- Enables process automation and workflow management
- Improve data quality (timeliness, coherence/consistency)
- → Document as you Survey (DaYS)

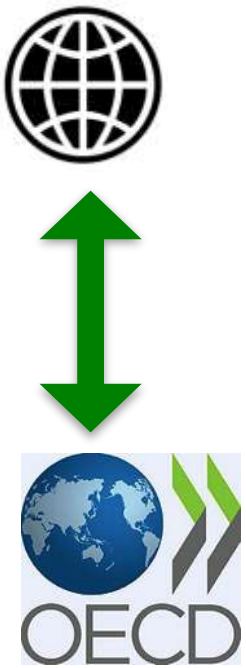
DDI for Longitudinal Studies



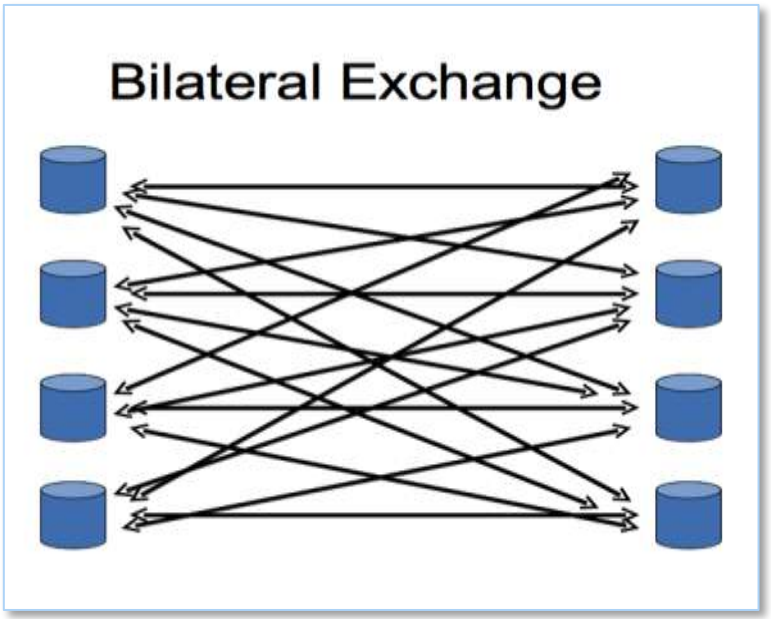
- DDI-Lifecycle allows metadata to be harmonized across waves
- Sharing metadata across survey cycles means less expensive survey development costs
- Researchers can find comparable data
- Leverage reuse, grouping, banks, common metadata, data element, etc.

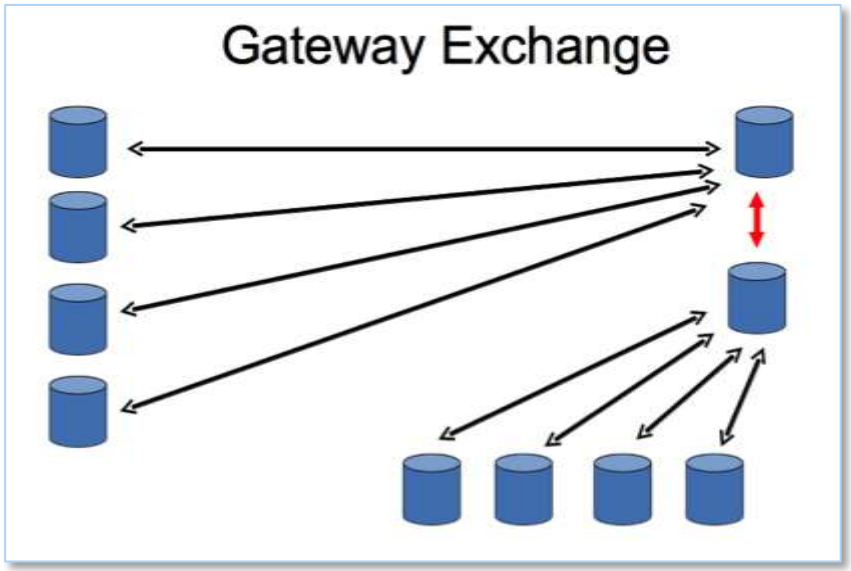
SDMX Exchange with Peers

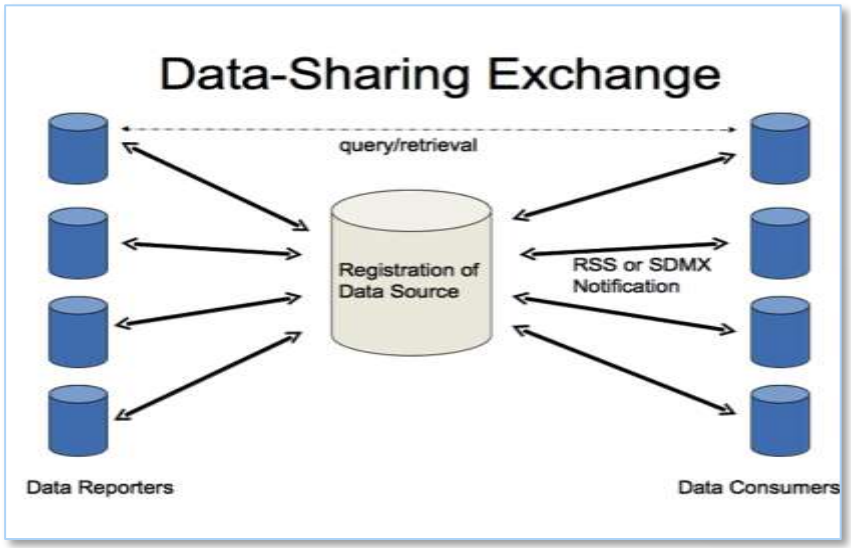
- SDMX-EDI and SDMX-ML are both able to exchange databases between peer organizations
- Structural metadata is also exchanged and can be read by counterparty systems
- Incremental updating is possible
- Increases degree of automation for exchange – lowers degree of bilateral, verbal agreement
- Can use “pull” instead of “push” if registry is deployed



Exchange Patterns







Which pattern is your organization using?

SDMX within an Organization

- **SDMX standard formats are also useful within an organization**
  - Many organizations have several disparate databases
  - Differences in database structure and content can make it difficult to use other system's data
  - SDMX-ML provides a way to loosely couple such databases, while facilitating exchange
  - An SDMX registry can allow visibility into other databases, while not affecting control or ownership of data

"What we have" checklist

- ✓ We have domain metadata standards and guidelines
- ✓ We have technology stack
- ✓ We have tools
- ✓ We have the expression of interest
- ✓ We "should" be in good shape!

CASIC

2012 Federal CASIC Workshops

Computer Assisted Survey Information Collection

Survey Uses of Metadata

Metadata: To Boldly Go Where No One Has Gone Before? PART3: Trouble with the Tribbles

Pascal Heus
pascal.heus@metadatatechnology.com

Vice-President
Metadata Technology North America
http://www.mtna.us

Executive Manager
Open Data Foundation
http://www.odaf.org

Progress and challenges

- Why is it not happening (faster)?
- What are the challenges / barriers?
- What are the day to day issues?

A Metadata Poor World

1	1	4	5	13
1	1	4	5	7
1	1	4	5	4
1	1	4	5	21
1	1	4	2	7
1	1	3	4	4
1	1	4	5	6
1	1	1	5	4
1	1	2	5	1
3	1	1	3	1
3	1	9	3	16
3	1	9	2	4
3	1	9	9	19
3	3	2	9	4
3	1	9	3	99

A Metadata Friendly World

Variable BRTCIT : Citizenship

Literal Question
"Are you... a British National (Overseas), a Full British Citizen - citizenship granted in the UK or a Full British Citizen - citizenship granted in Hong Kong?"

Categories	Value	N	
British National Overseas	1	11	9.4%
Full British Citizen	2	72	61.5%
Full Brit Citizen granted in Hong Kong	3	27	23.1%
Other, Don't know	4	7	6.0%
Does not apply	-9	140249	
No answer	-8	0	

Summary statistics

Valid cases	117
Minimum	1
Maximum	4
Mean	2.25641

This variable is numeric

Universe
Applies: respondent is a British National who was born in Hong Kong or China.

Total Responses
Summation of listed categories: 140366

What about statistical data?

- Do we live in a metadata poor statistical data world?
- Yes and no. It's not that bad, but it's not that great
  - Microdata: software rarely goes beyond the data dictionary
  - Aggregated data: HTML, excel
  - Documentation: PDF, Word
- Data (micro of macro) is often produced and disseminated with little metadata
- We don't have a good inventory (metadata) and lack intelligent file management systems (data/docs)
- This is where we must start

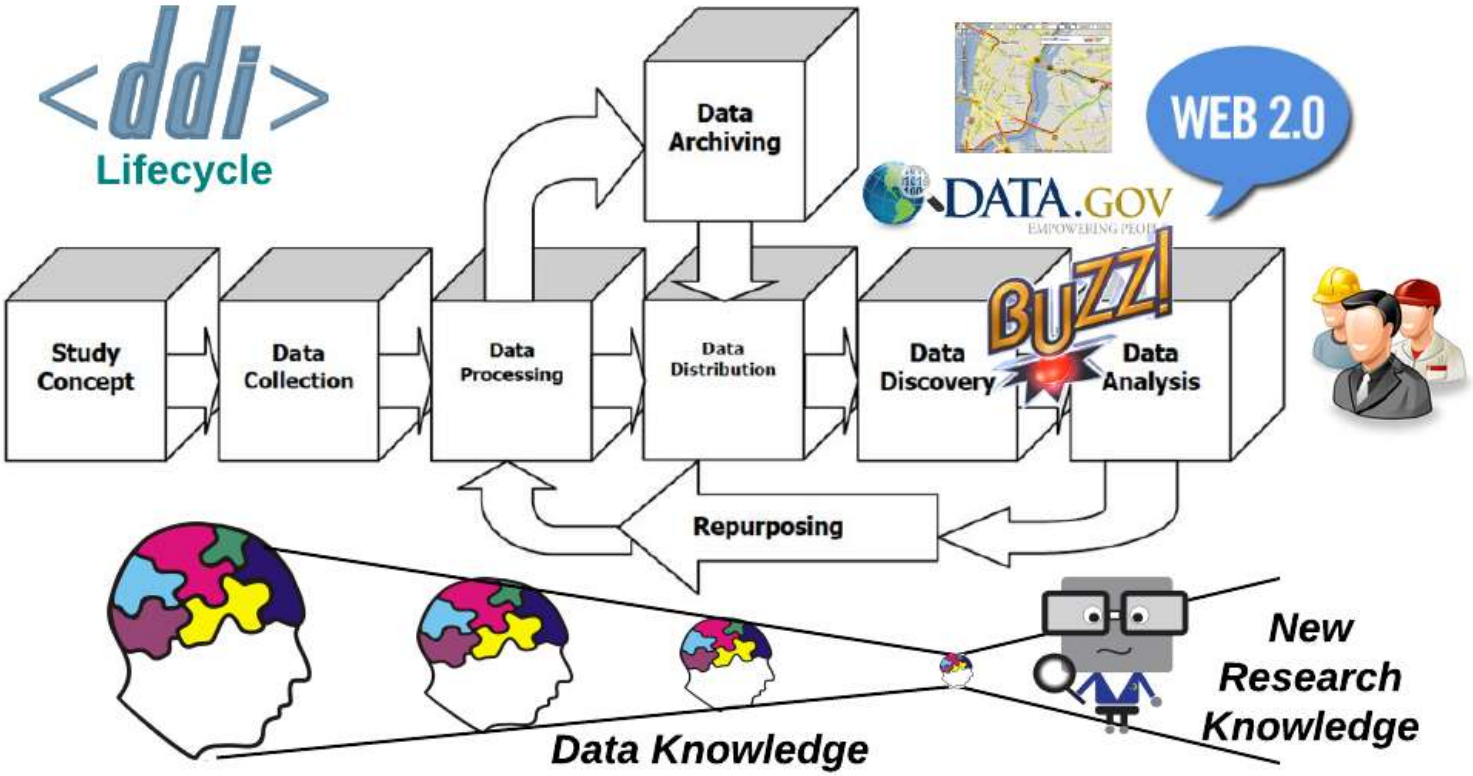
What to do?

- Is this just us?
  - No, these issues are universal and not domain specific
  - The rise of the Internet has prompted industry to take action (B2B, B2C, eCommerce)
  - IT technology and standards have emerged to solve this
  - But the statistical world has been slow to adopt
- Solution?
  - Simple in theory: deploy management framework, train/educate, manage change and integration/migrartion
- How?
  - Agree on format: use standards such as DDI, SDMX, ISO 11179, and the like (for communicating with others)
  - Leverage technology: XML
  - Change practices: it's not just a technical challenge

Why aren't we doing it?

- We don't know enough about it
  - Statistical agencies are not IT experts
  - → Need to better inform stakeholders, managers, users
- We don't like change
  - Traditions are strong and the agencies mandate is focused on data
  - → change management, executive support, non-intrusive strategies
- The tools we typically use are ill-equipped
  - → complement with metadata driven tools and pressure vendors for better tools
- How much does it costs?
  - → Minimal compare to the effort going into producing data
  - → Significant saving down the road (automation, quality, reduce burden)
- Does it work?
  - → yes, but we need more innovators, early adopters, champions
  - → the Internet is a pretty good success story

Importance of early capture



- There is pressure to deliver to end users
- But deep knowledge resides with the producer
- Delivering quality metadata requires early stage capture → this is where we have to focus first!

Importance of core metadata

- We do not pay enough attention to fundamental metadata
  - Classifications, Concepts, Universes, Questions (banks)
- There seem to be an implicit belief that data is linkable
- These are absolutely necessary to achieve
  - harmonization and linkages
  - Effective search and discovery for large collections

Official country names used by the ISO 3166/MA	Numeric	Alpha-3	Alpha-2
Afghanistan	004	AFG	AF
Åland Islands	248	ALA	AX
Albania	008	ALB	AL
Algeria	012	DZA	DZ
American Samoa	016	ASM	AS
Andorra	020	AND	AD
Angola	024	AGO	AO
Anguilla	660	AIA	AI
Antarctica	010	ATA	AQ
Antigua and Barbuda	028	ATG	AG
Argentina	032	ARG	AR
Armenia	051	ARM	AM
Aruba	533	ABW	AW
			AU
			AT
			AZ
			BS
			BH

Funding for metadata and tools

- Billions are committed every year for data production
  - US 2012: ~$4 billion (~$800M census)
  - Metadata are often poorly budgeted (if at all)
  - Data production, preservation, discovery, access, and analysis infrastructure do not sufficiently leverage today's Information Technology (impact costs, quality)
- If you were publishing a book, what % of your budget would go into packaging, distribution, marketing, etc.?
- Data production can be expensive, but a higher % of the budget must go into supporting quality and broad usage (access, usability, quality, linking, etc.)

"Tribbles" checklist

- Things won't change on their own
- We cannot deliver what is being promised without sound data/metadata management framework that supports automation
- We need to start at the core of the problem
- We need to properly and effectively invest in metadata
- Technology issues can be solved
- Change management and non-disruptive transition strategies are essential
- So we "could" be in good shape, but need sound action plan

CASIC

2012 Federal CASIC Workshops

Computer Assisted Survey Information Collection

Survey Uses of Metadata

Metadata: To Boldly Go Where No One Has Gone Before? PART4: Next Generation?

Pascal Heus
pascal.heus@metadatatechnology.com

Vice-President
Metadata Technology North America
http://www.mtna.us

Executive Manager
Open Data Foundation
http://www.odaf.org

Need more motivations?

- Transitioning into Next Generation environment
  - Demand for data has dramatically increased
  - Nature and amount of data have changed
  - Need to address Accessibility, Timeliness, Consistency, Linking/Harmonization
- Globalization
  - We need to have a big picture of the world (population growth, economy, …)
  - Cross-agency collaboration and exchange are no longer an option
  - We cannot work is isolation
  - This requires common "language"
- Solve challenge of balancing openness and privacy
- Need to capture knowledge in a digital world
- Need to reduce burden and costs
- Make everyone's life easier

Beyond the Data: Impact

- We also have to look beyond metadata and technology, remember the big picture
- These are our instruments to measure the health and state of nations
- Data from respondent is ultimately for decision makers (evidence based policies, monitoring)
- Impact is on people, societies, living conditions
- Current population dynamics and world economy requires global knowledge
- There is a sense of urgency…

Grand Unification?

- Sound information management framework = business processes + standards + technology?
- GSBPM/GSIM + SDMX/DDI + XML/RDF + SOA + DBMS + IT
- But need to complement with:
  - Change management, executive support
  - Training / Education, Knowledge sharing
  - Non-disruptive integration/migration strategy (business continuity)
- Roadmap
  - If we want to grant users and executives wishes, we must start with fundamental data management issues

Conclusions

- We have a beautiful list of objectives
  …can't be achieved without sound metadata and high quality data
- We have the standards, expertise, and technology
  ….but not the content
- We have funding
  …but not enough is directed towards solving the issues
- Pioneers and innovators have emerged
  …but not enough of them, particularly in the US
- We can develop tools and guidelines
  …but changing the way we work may be the hardest challenge (the human factor)
- So change is/could be on the way
  …but requires concerted and coordinated efforts
- Starting today?

IASSIST 2012

CASIC
2012 Federal CASIC Workshops
Computer Assisted Survey Information Collection

Data Science for a Connected World: Unlocking and Harnessing the Power of Information

June 4-8
Marvin Centre, George Washington University
http://www.iassist2012.org/