

what can
**META
DATA**
do for

you



Metadata Repository Workshop
January 23rd 2012

**Using Metadata in Data
Archiving, Data Analysis and
Study Evaluation**

Mr. Jeremy Iverson
Partner
Algenta Technologies

Dr. Timothy Mulcahy
Project Director
Data Enclave
NORC at University of
Chicago

Dr. William Block
Director
Cornell Institute for Social
and Economic Research
(CISER)

Mr. Pascal Heus
Executive Manager
Open Data Foundation
Vice-President
Metadata Technology
North America

what can
**META
DATA**
do for

you



Metadata Repository Workshop
January 23rd 2012

**Metadata:
What is it all about?**



What is metadata?



DATA ABOUT DATA



- We need a better definition...
- Metadata is information we use every day to make decisions, navigate in our environment, share knowledge, search, and learn about things
- Metadata makes our life easier

What is Metadata?

It's not just 'data about data' ...



| 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 + Trans / Trans | 4 g 21 % 0.2 g |
| 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...

Everyday Metadata

“Human Readable” Metadata

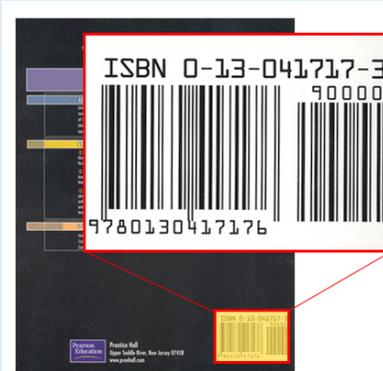
“Machine-actionable” Metadata



| S-Bahn | | Gleis | |
|--------|-----------------------|-------------------------|-------|
| Linie | Ablfahrt | | |
| S3 | 14.33 Stadelhofen | Effretikon | 23 |
| S3 | 14.34 Altstetten | Affoltern a/A | 21 |
| S9 | 14.34 Seinau | Triemli Wilikon Waldegg | 2 |
| S10 | 14.37 Oerlikon | Oberglatt | 22 |
| S5 | 14.37 Enge Thalwil | Langnau-G. | 53 |
| S2 | 14.38 Seinau Adliswil | Meilen Stäfa | 1 |
| S4 | 14.38 Stadelhofen | Uster Wetzikon | 23/24 |
| S7 | 14.42 Oerlikon | Dietikon | 54 |
| | | präflikon SZ | 21/22 |
| | | | 52 |
| | | | 23/24 |



| Nutrition Facts | | Valeur nutritive | |
|---|--------|----------------------|--|
| Per 1 bowl (300 g) / Pour 1 bol (300 g) | | | |
| Amount | | % Daily Value | |
| Teneur | | % valeur quotidienne | |
| Calories / Calories | 440 | | |
| Fat / Lipides | 19 g | 29 % | |
| Saturated / Saturés | 4 g | 21 % | |
| + Trans / Trans | 0.2 g | | |
| 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 % | |





Everyday Metadata



YAHOO!



Weather Forecast

Compact | Classic | Full

1°C
Mostly Cloudy

Location: Washington, DC
Mostly Cloudy

| Location | Today | Tomorrow | Tuesday |
|----------------|---------|----------|---------|
| Washington, DC | 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 ago
- 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

| Team | Score | Status |
|------------|-------|--------|
| Washington | 20 | Final |
| Seattle | 17 | |

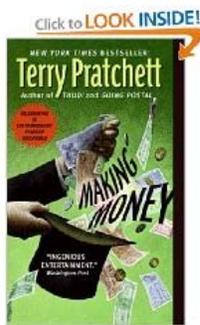
Currency Converter

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

| Currency | U.S. \$ | Yen | Euro | U.K. £ |
|-------------|---------|----------|--------|--------|
| 1 U.S. \$ = | 1 | 95.05 | 0.7936 | 0.1 |
| 1 Yen = | 0.0105 | 1 | 0.0083 | 0.1 |
| 1 Euro = | 1.2601 | 119.7725 | 1 | 0.1 |
| 1 U.K. £ = | 1.4891 | 141.5389 | 1.1817 | 1 |

Quotes from Yahoo!

| Stock | Price | Change | % Change |
|-------|--------|--------|----------|
| 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% |



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**



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 | -9140249 | | |
| 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
- What can we improve on this?



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: capture more and better metadata
- **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**
 - and the mandate is focused on data
 - → change management, executive support, non-intrusive strategies
- **The tools we use are not well equipped**
 - → complement with metadata driven tools and pressure vendors for better tools
- **How much does it cost?**
 - → 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
 - → but the Internet is a pretty good success story



More reasons...



- **Demand has changed**
 - Demand for data has dramatically increased
 - Amount data is dramatically increasing
- **Globalization**
 - We need have a big picture of the world
 - Collaboration and exchange are necessary
- **Transparency**
 - Data.gov
- **Reduce burden and costs**
- **Preserve knowledge in a digital world**
 - Need to store in non-proprietary formats (ASCII is not enough)

what can
**META
DATA**
do for

you



Metadata Repository Workshop
January 23rd 2012

**Leveraging Metadata
using XML & DDI**



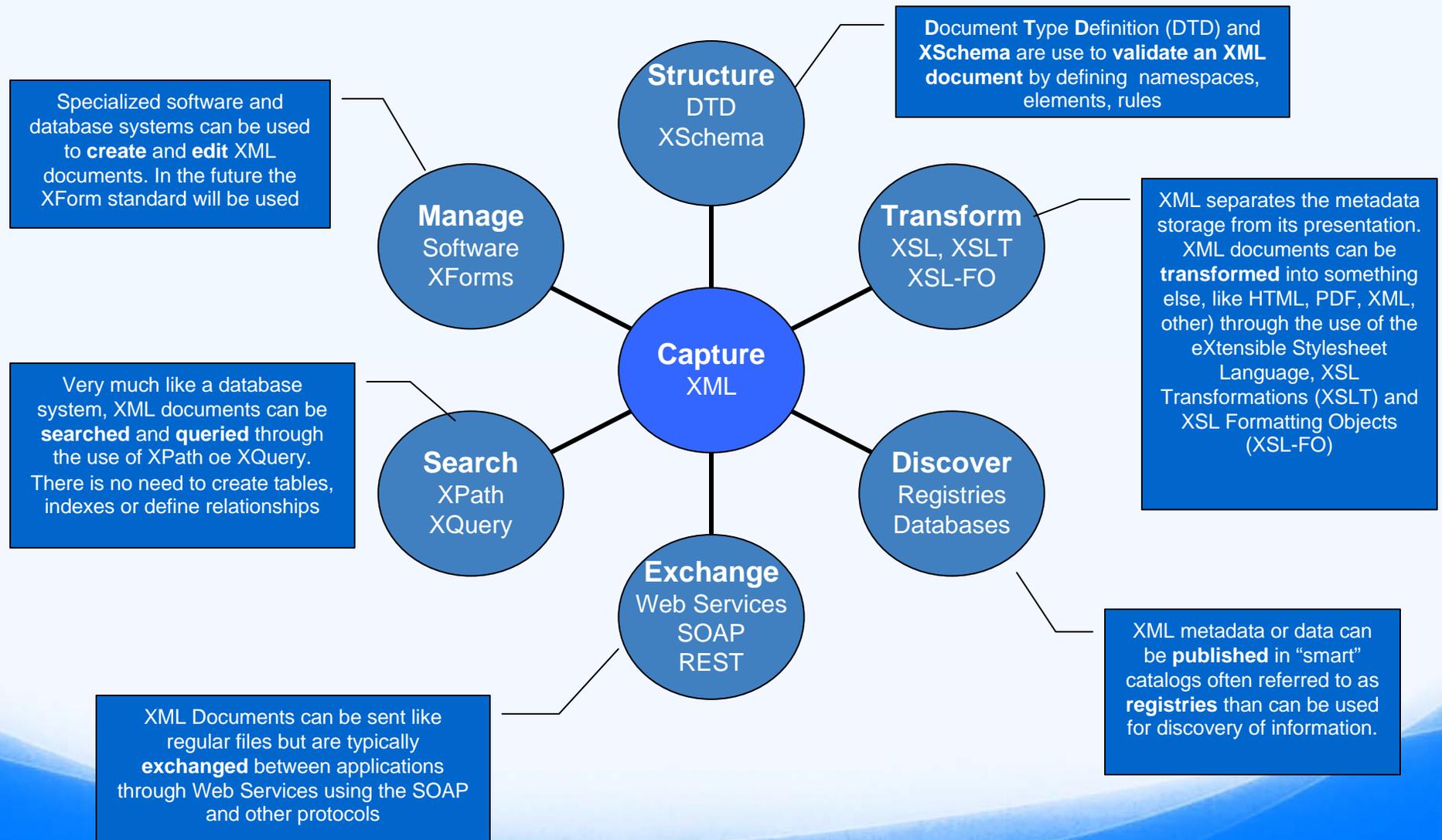
What is XML?



- Today's Universal language on the web
- Purpose is to facilitate sharing of structured information across information systems in a generic fashion
- XML stands for eXtensible Markup Language
 - eXtensible → can be customized
 - Markup → tags, marks, attach attributes to things
 - Language → syntax (grammatical rules)
- **HTML (HyperText Markup Language) is a markup language but not extensible! It is also concerned about presentation, not content.**
- XML is a text format (not a binary black box)
- XML is also a collection of technologies (built on the XML language)
- It is platform independent and is understood by modern programming languages (C++, Java, .NET, PHP, perl, etc.)
- It is both machine and human readable



XML is a set of technologies





Data Documentation Initiative



- XML specification focusing on microdata
- 1.0 published in 2000
 - emerged from the data archive community (ICPSR)
- Governed by DDI Alliance
 - 35+ members
- Captures metadata on
 - Survey, files, variables, value labels, summary statistics
 - But also concepts, universes, questions, geography, provenance, access policies, and more
 - Wrap comprehensive knowledge around data
- Two flavors: DDI-Codebook and DDI-Lifecycle



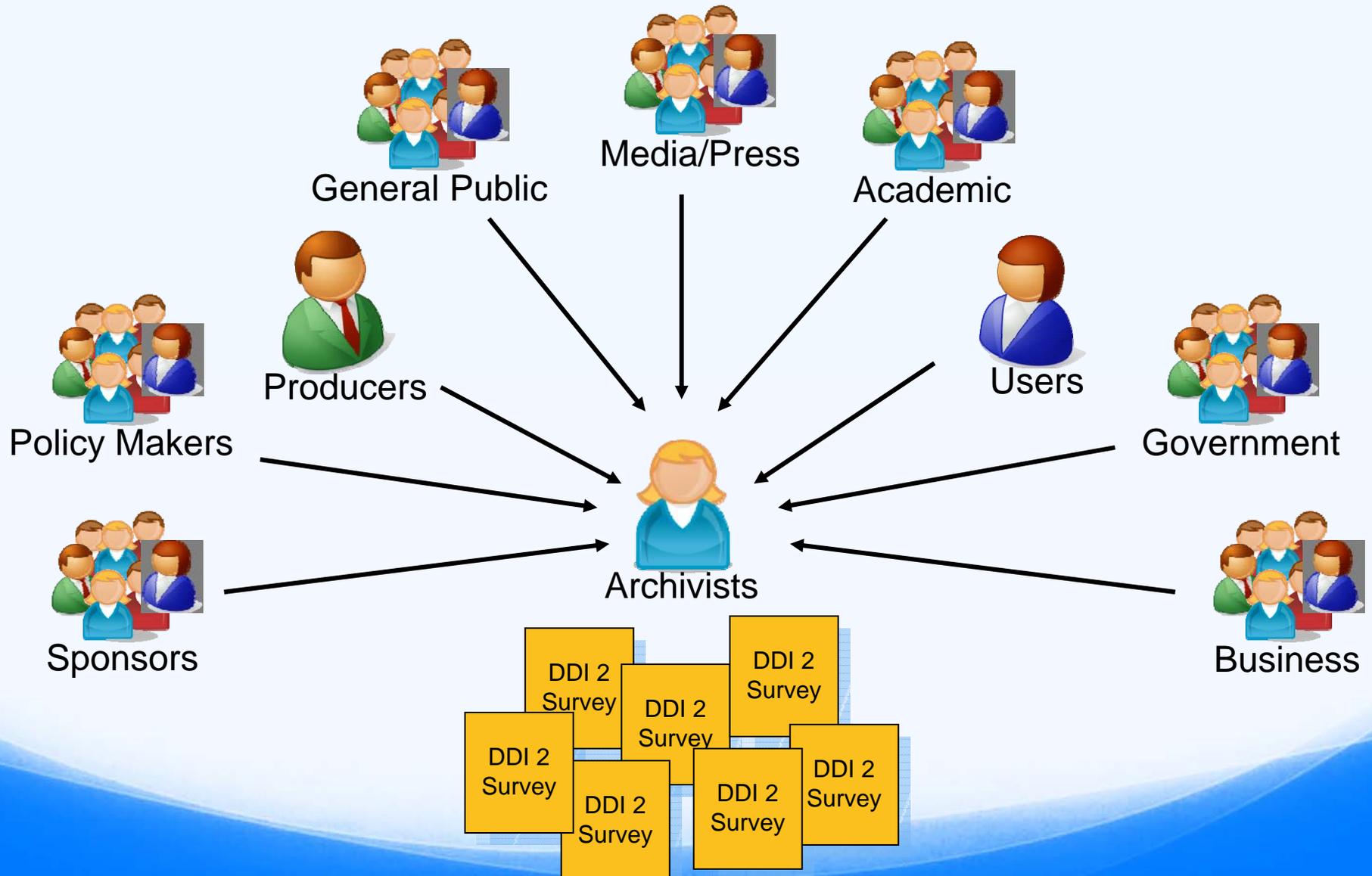
DDI Flavors



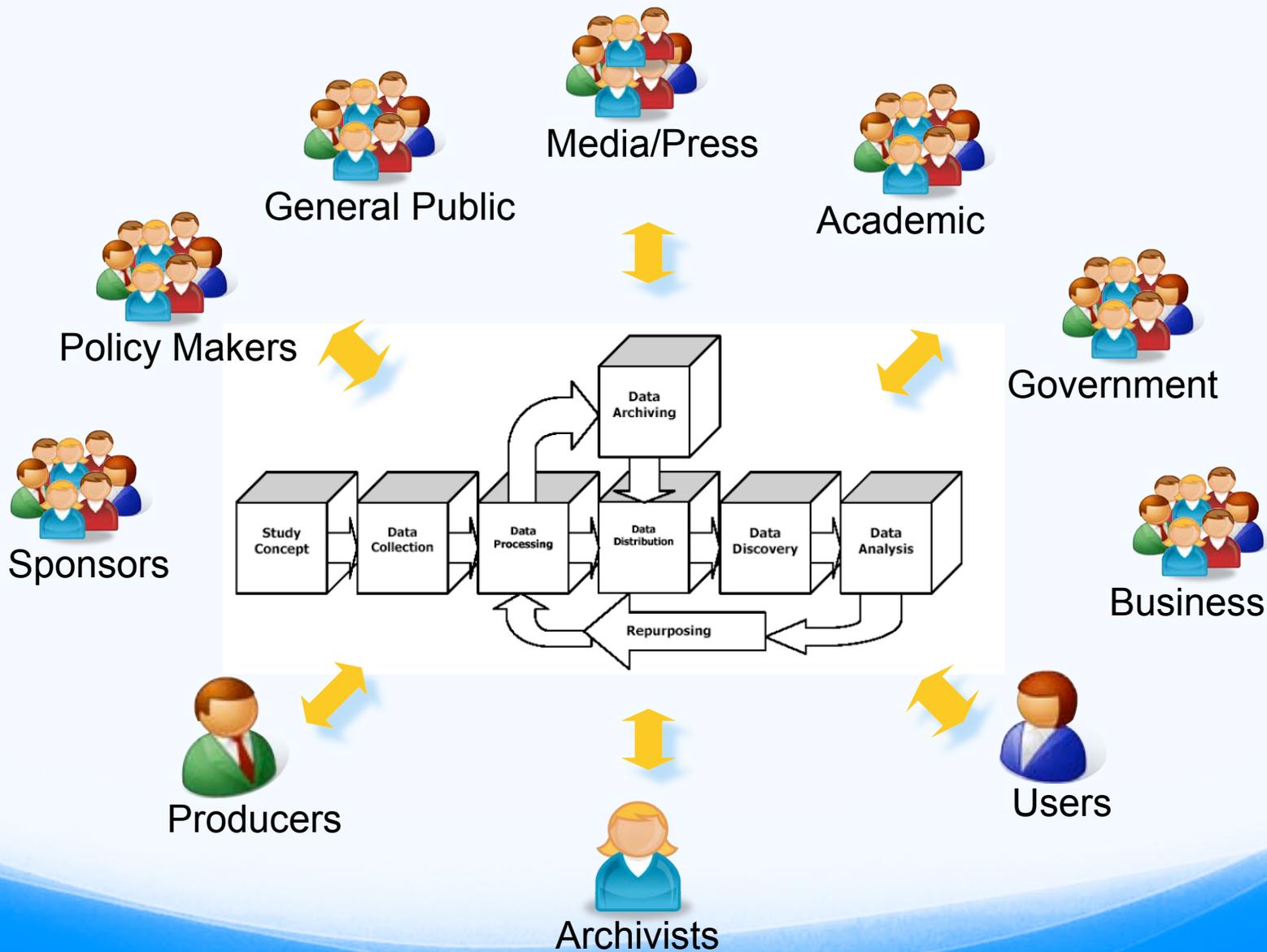
- **DDI-Codebook (v1.0 - v2.5)**
 - Archive/Dissemination
 - Pros: Mature, simple to use, widely adopted, tools available
 - Cons: focus on documenting single survey instance, limited use cases, single maintainer, no metadata reuse, required variables to exist
- **DDI-Lifecycle (v3.0 – v3.1)**
 - Supports entire data lifecycle
 - Pros: high level of reuse, many maintainers, work across surveys, numerous applications / use cases, more formal model
 - Cons: more complex to use, tools emerging, requires IT and technical capacity



DDI-Codebook Perspective



DDI-Lifecycle perspective





Example of metadata



Module/Concepts
Universe
Instruction

| EDUCATION MODULE | | | | | | | | | | | | | | | | | |
|---|---|--|-----------|--|---|---|-----|---|------|--|--|---|---|---|-----------|---|--|
| <i>If interview takes place between two school years, use alternative wording found in Appendix 1.</i> | | | | | | | | | | | | | | | | | |
| <i>For persons age 5 or over ask Qs. 15 and 16</i> | | | | | <i>For children age 5 through 17 years, continue on, asking Qs. 17-22</i> | | | | | | | | | | | | |
| 14. <i>Line no.</i> | | 15. HAS (<i>name</i>) EVER ATTENDED SCHOOL? | | 16. WHAT IS THE HIGHEST LEVEL OF SCHOOL (<i>name</i>) ATTENDED? WHAT IS THE HIGHEST GRADE (<i>name</i>) COMPLETED AT THIS LEVEL? LEVEL: 1 PRIMARY 2 SECONDARY 3 HIGHER 4 NON-STANDARD CURRICULUM 9 DK GRADE: 99 DK <i>If less than 1 grade, enter 00.</i> | | 17. IS (<i>name</i>) CURRENTLY ATTENDING SCHOOL? | | 18. DURING THE CURRENT SCHOOL YEAR, DID (<i>name</i>) ATTEND SCHOOL AT ANY TIME? | | 19. SINCE LAST (<i>day of the week</i>), HOW MANY DAYS DID (<i>name</i>) ATTEND SCHOOL? | | 20. WHICH LEVEL AND GRADE IS/ <i>WAS</i> (<i>name</i>) ATTENDING? LEVEL: 1 PRESCHOOL 2 PRIMARY 3 SECONDARY 4 NON-STANDARD CURRICULUM 9 DK GRADE: 99 DK | | 21. DID (<i>name</i>) ATTEND SCHOOL LAST YEAR? | | 22. WHICH LEVEL AND GRADE DID (<i>name</i>) ATTEND LAST YEAR? LEVEL: 1 PRESCHOOL 2 PRIMARY 3 SECONDARY 4 NON-STANDARD CURRICULUM 9 DK GRADE: 99 DK | |
| | | 1 YES ⇨ Q.18 2 NO ⇨ NEXT LINE | | | 1 YES ⇨ Q.19 2 NO | | | 1 YES 2 NO ⇨ Q.21 | | | <i>Insert number of days in space below.</i> | | | 1 YES 2 NO ⇨ NEXT LINE 9 DK ⇨ NEXT LINE | | | |
| | | Questions | | Classifications (some reusable) | | | | | | | | | | | | | |
| LINE | Y | NO | LEVEL | GRADE | YES | NO | YES | NO | DAYS | LEVEL | GRADE | Y | N | DK | LEVEL | GRADE | |
| 01 | 1 | 2 ⇨ NEXT LINE | 1 2 3 4 9 | ___ | 1 | 2 | 1 | 2 | ___ | 1 2 3 4 9 | ___ | 1 | 2 | 9 | 1 2 3 4 9 | ___ | |
| 02 | 1 | 2 ⇨ NEXT LINE | 1 2 3 4 9 | ___ | 1 | 2 | 1 | 2 | ___ | 1 2 3 4 9 | ___ | 1 | 2 | 9 | 1 2 3 4 9 | ___ | |
| 03 | 1 | 2 ⇨ NEXT LINE | 1 2 3 4 9 | ___ | 1 | 2 | 1 | 2 | ___ | 1 2 3 4 9 | ___ | 1 | 2 | 9 | 1 2 3 4 9 | ___ | |
| 04 | 1 | 2 ⇨ NEXT LINE | 1 2 3 4 9 | ___ | 1 | 2 | 1 | 2 | ___ | 1 2 3 4 9 | ___ | 1 | 2 | 9 | 1 2 3 4 9 | ___ | |
| 05 | 1 | 2 ⇨ NEXT LINE | 1 2 3 4 9 | ___ | 1 | 2 | 1 | 2 | ___ | 1 2 3 4 9 | ___ | 1 | 2 | 9 | 1 2 3 4 9 | ___ | |
| 06 | 1 | 2 ⇨ NEXT LINE | 1 2 3 4 9 | ___ | 1 | 2 | 1 | 2 | ___ | 1 2 3 4 9 | ___ | 1 | 2 | 9 | 1 2 3 4 9 | ___ | |
| 07 | 1 | 2 ⇨ NEXT LINE | 1 2 3 4 9 | ___ | 1 | 2 | 1 | 2 | ___ | 1 2 3 4 9 | ___ | 1 | 2 | 9 | 1 2 3 4 9 | ___ | |
| <i>Now for each woman age 15-49 years, write her name and line number at the top of each page in the Women's Questionnaire.</i> | | | | | | | | | | | | | | | | | |
| <i>For each child under age 5, write his/her name and line number AND the line number of his/her mother or caretaker at the top of each page in the Children's Questionnaire.</i> | | | | | | | | | | | | | | | | | |
| <i>You should now have a separate questionnaire for each eligible woman and child in the household.</i> | | | | | | | | | | | | | | | | | |

Value level Instruction (skip)

Instruction



Common metadata example



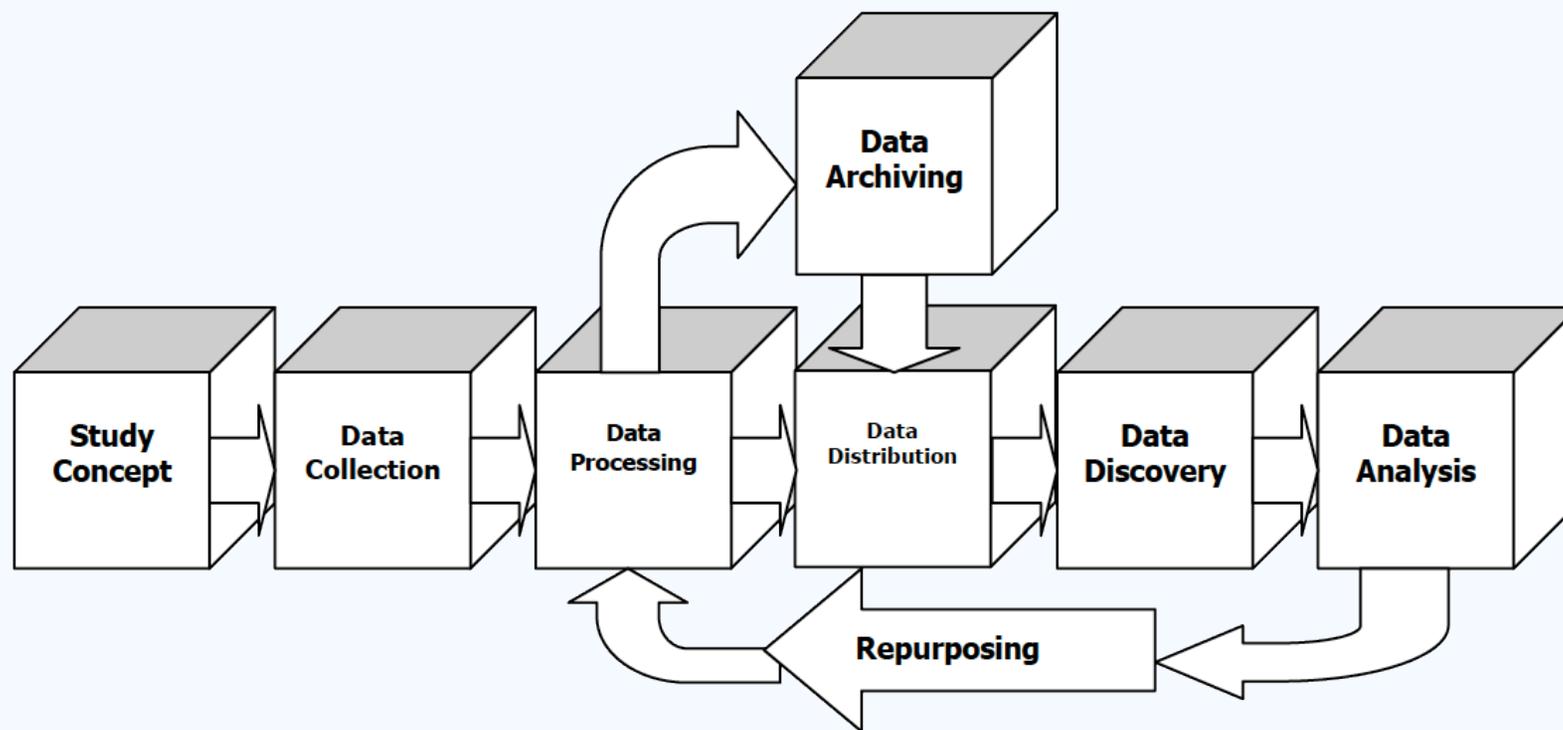
| Official country names used by the ISO 3166/MA | Numeric | Alpha-3 | Alpha-2 |
|--|---------|---------|---------|
| Afghanistan | 004 | AFG | AF |
| Åland Islands | 248 | ALA | AX |
| Albania | 008 | ALB | AL |
| Algeria | 012 | DZA | DZ |
| American Samoa | 016 | ASM | AS |
| Andorra | 020 | AND | 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 |
| Australia | 000 | AUS | AU |
| Austria | 009 | AUT | AT |
| Azerbaijan | 031 | AZE | AZ |
| Bahamas | 064 | BHS | BS |
| Bahrain | 067 | BHR | BH |



| Neoplasms (C00-D48) | |
|-------------------------|--|
| C00-C97 | Malignant neoplasms |
| C00-C75 | Malignant neoplasms, stated or presumed to be primary, of specified sites, except of lymphoid, haematopoietic and related tissue |
| C00-C14 | Lip, oral cavity and pharynx |
| C15-C26 | Digestive organs |
| C30-C39 | Respiratory and intrathoracic organs |
| C40-C41 | Bone and articular cartilage |
| C43-C44 | Skin |
| C45-C49 | Mesothelial and soft tissue |
| C50 | Breast |
| C51-C58 | Female genital organs |
| C60-C63 | Male genital organs |
| C64-C68 | Urinary tract |
| C69-C72 | Eye, brain and other parts of central nervous system |
| C73-C75 | Thyroid and other endocrine glands |
| C76-C80 | Malignant neoplasms of ill-defined, secondary and unspecified sites |
| C81-C96 | Malignant neoplasms, stated or presumed to be primary, of lymphoid, haematopoietic and related tissue |
| C97 | Malignant neoplasms of independent (primary) multiple sites |
| D00-D09 | In situ neoplasms |
| D10-D36 | Benign neoplasms |
| D37-D48 | Neoplasms of uncertain or unknown behaviour [see note before D37] |



DDI Lifecycle

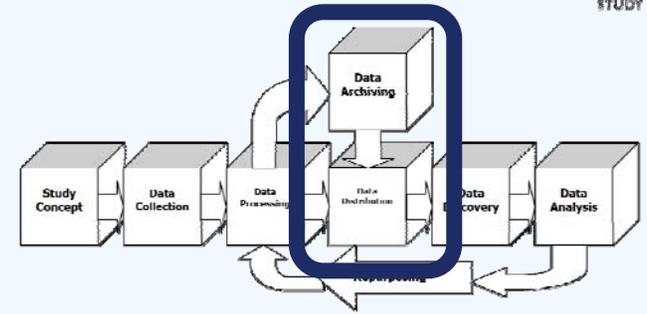




DDI for Archive/Preservation

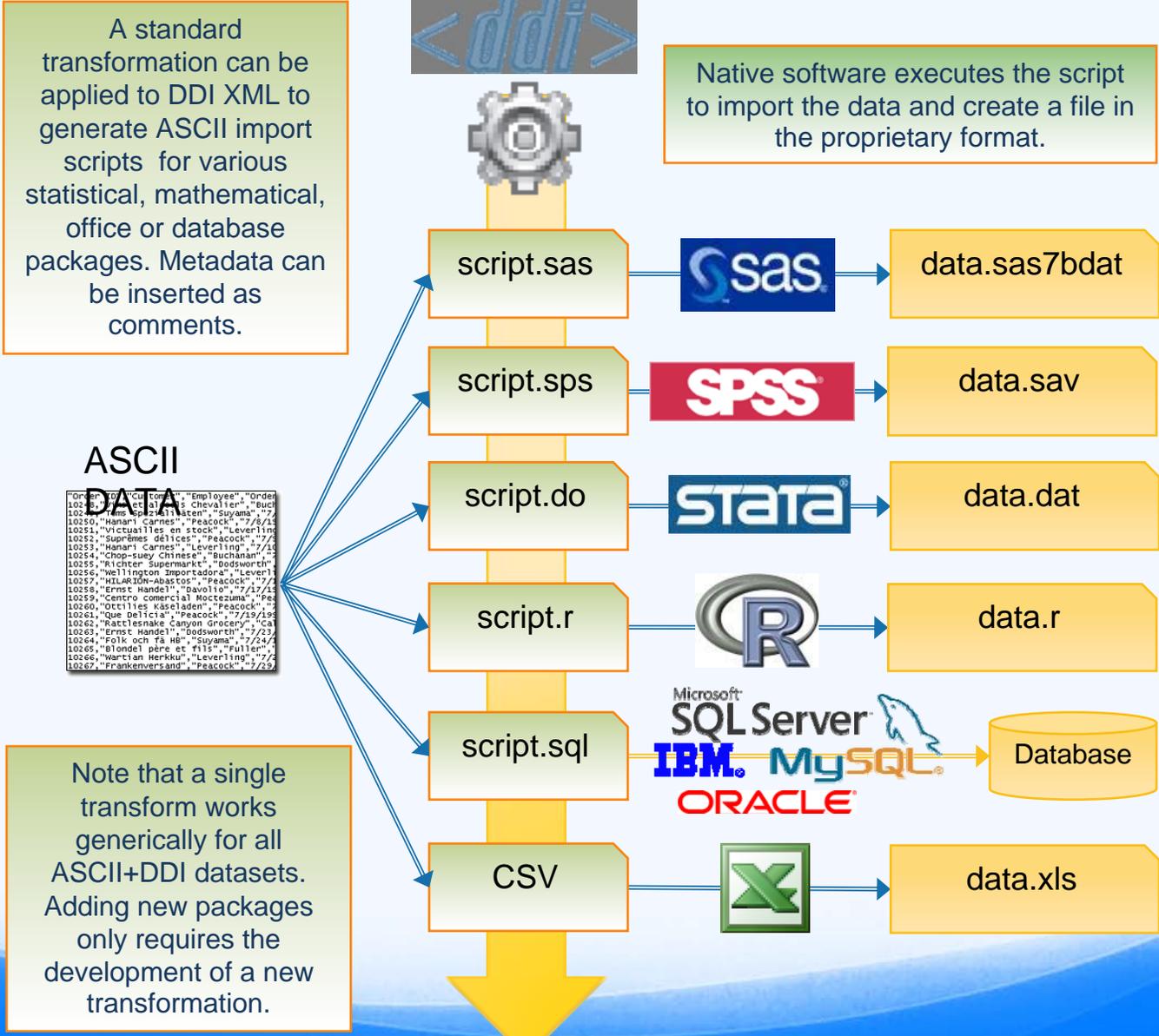


- Captures comprehensive information about surveys and their data
- DDI is widely used by national statistical agencies, data archives, research centers around the globe
- ASCII + DDI is also a powerful combination for long term preservation (non-proprietary text format)





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.

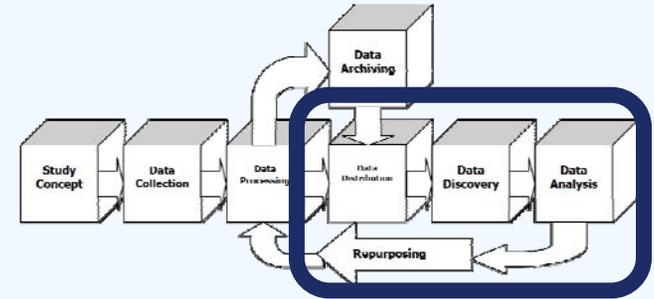
Note that a single transform works generically for all ASCII+DDI datasets. Adding new packages only requires the development of a new transformation.



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

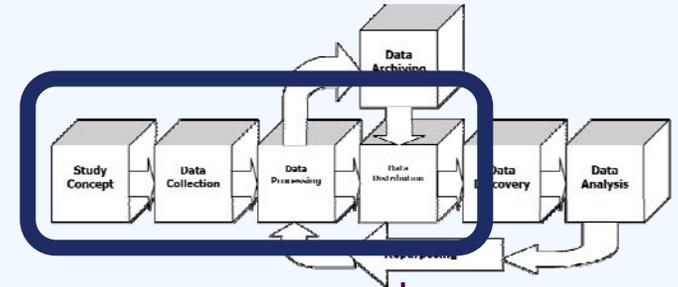




DDI for Production

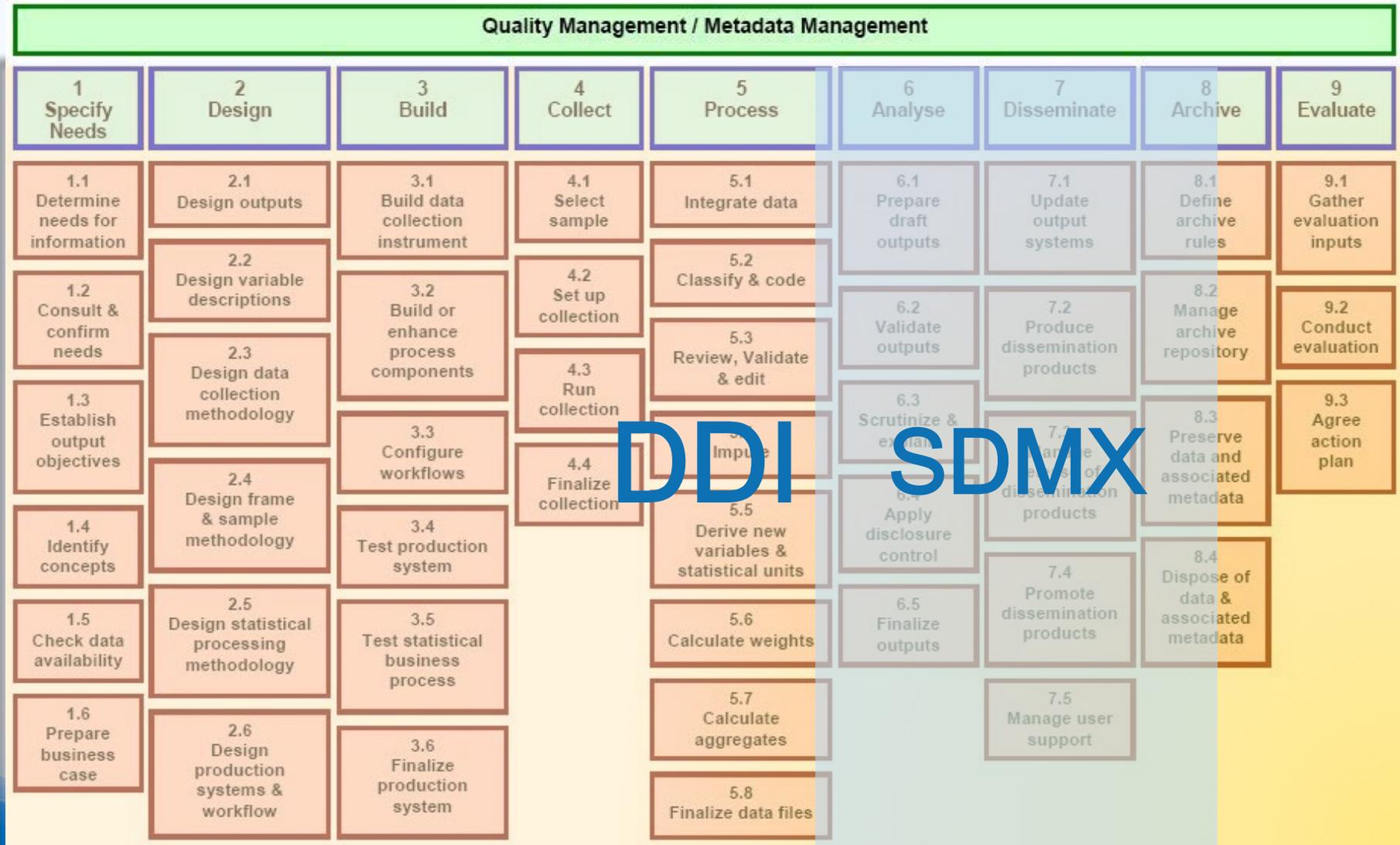


- Designed for use from day 1 of a study or program
- Manage common metadata elements - such as concepts, universes, geography – across surveys or waves, or even agencies
- Supports classifications, question, variable, concept banks
- Enables process automation and workflow management
- Improve data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





DDI and GSBPM



DDI SDMX

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

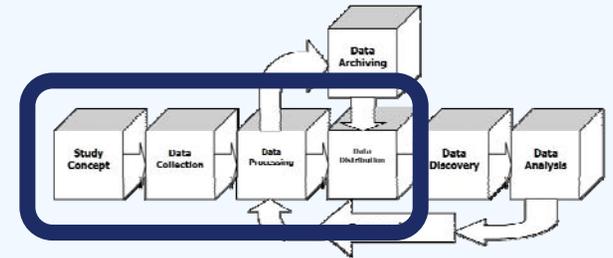
NCS Metadata Repository Workshop – Jan 23rd 2012



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.





NIH Support for DDI Tools



- **SBIR Phase 1: Automated Survey Instrument Documentation, Analysis, and Testing Software**
 - October 2006 – August 2007
 - 1 R43 AG027612-01
- **Applying the Data Documentation Initiative (DDI) to MIDUS**
 - September 2009 – August 2011
 - 5 RO3 AG032268-02
- **SBIR Phase 1: Open Standards-Based Data Extraction Web Tool for Complex Longitudinal Datasets**
 - April 2011 – November 2011
 - 1 R43 AG039898-01



Summary



- Industry standard IT technologies are available to support the management and exchange of metadata (public or private)
- Domain specific mature and powerful XML specifications are available for socio-economic data / official statistics
- A complete solutions often combines several standards designed to inter-operate.
- The Data Documentation Initiative (DDI) is the recommended specification for “microdata”
- The Statistical Data Exchange Standard (SDMX) is the recommended standard for aggregated data / statistics
- Adoption across the US statistical system would greatly benefit all stakeholders

what can
**META
DATA**
do for

you



Metadata Repository Workshop
January 23rd 2012

DDI Stories



NSF-Census Research Network (NCRN)



- **Cornell NSF-Census Research Node: Integrated Research, Support, Training, and Data Documentation**
- **1 of 8 Nodes funded in the Network (\$1.2M – \$3M each)**
- **Investigators: John Abowd, William Block, Lars Vilhuber, and Ping Li**
- **The Comprehensive Census Bureau Metadata Repository (CCBMR)**
- **Socio-economic / official statistics often have need for confidentiality restrictions/privacy.**
- **Similar to health data (and thus relevant to National Children's Study)**



The Death Knell for Public-use Data

- Sounded by young scholars pursuing research programs that mandate inherently identifiable data: geospatial relations, exact genome data, networks of all sorts, linking administrative records.
- These researchers acquire authorized restricted access to the confidential identifiable data and perform their analyses in secure environments.
- But they don't leave behind the scientific trail that has made public-use files so important.



The Comprehensive Census Bureau Metadata Repository (CCBMR)

- Facilitates access to detailed metadata on
 - Restricted-access data from outside an RDC while enabling fine-grained control over confidential information for the (Longitudinal Business Database (LBD), American Community Survey (ACS), American Housing Survey (AHS), Longitudinal Employer-Household Dynamics (LEHD))
 - Public-use datasets inside restricted-access areas (IPUMS, CPS)
- Expands the notion of metadata to include user-generated components (notes, programs, *etc.*)



Structure of CCBMR Metadata Schema

- Based as much as possible on existing schemas: *e.g.*, DDI, SDMX, and DataCite
- Modifies or adds fields/elements and/or attributes as necessary
- Example (DDI-based):

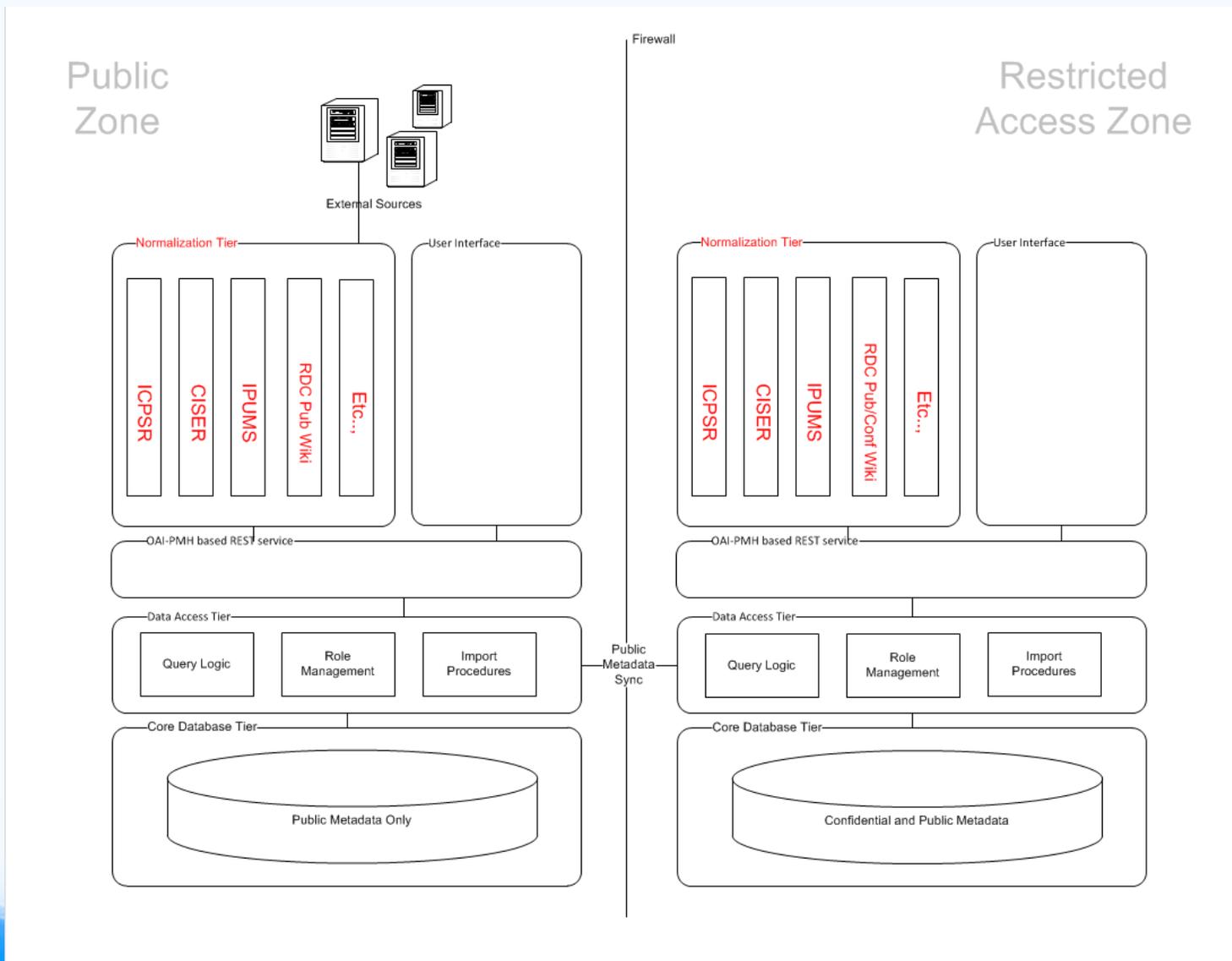
RDC Metadata (complete)

```
<d:VariableSet>
  <d: VariableItem> ...:<d:/VariableItem>
  <d:Disclosability>
    <d:min disclosable="yes">0</d:min>
    <d:max disclosable="no">345678</d:max>
  </d:Disclosability>
</d:VariableSet>
```

Derived Public Use Metadata (limited)

```
<d:VariableSet>
  <d: VariableItem> ...:<d:/VariableItem>
  <d:Disclosability>
    <d:min>0</d:min>
    <d:max>not disclosable</d:max>
  </d:Disclosability>
</d:VariableSet>
```

Draft of an Enterprise Application





Summary (compared to NCRN)



- Industry standard IT technologies are available to support the management and exchange of metadata (public or private) ✓
- Domain specific mature and powerful XML specifications are available for socio-economic data / official statistics ✓
- The complete solutions combines several standards designed to inter-operate. ✓
- The Data Documentation Initiative (DDI) is the recommended specification for “microdata” ✓
- The Statistical Data Exchange Standard (SDMX) is the recommended standard for aggregated data / statistics ✓
- Adoption across the US statistical system would greatly benefit all stakeholders ✓
- Seems very relevant to NCS needs (confidential and longitudinal data)



Other projects leveraging DDI



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



Longitudinal Surveys using DDI



- **Midlife in the United States (MIDUS)**
 - Data documentation web site powered by DDI- Lifecycle
<http://midus.colectica.org/>
- **Wisconsin Longitudinal Study (WLS)**
 - Uses DDI-Codebook to generate data dictionaries
 - Exploring DDI-Lifecycle to enable harmonization with HRS, MIDUS, SHARE, and other longitudinal studies
- **Abroad**
 - **United Kingdom - Birth Cohort Study**
 - 100K children, 2012-2020, £28M
 - **Germany - National Educational Panel Study (NEPS)**
 - 60K participants, 6 cohorts, 2007-2019

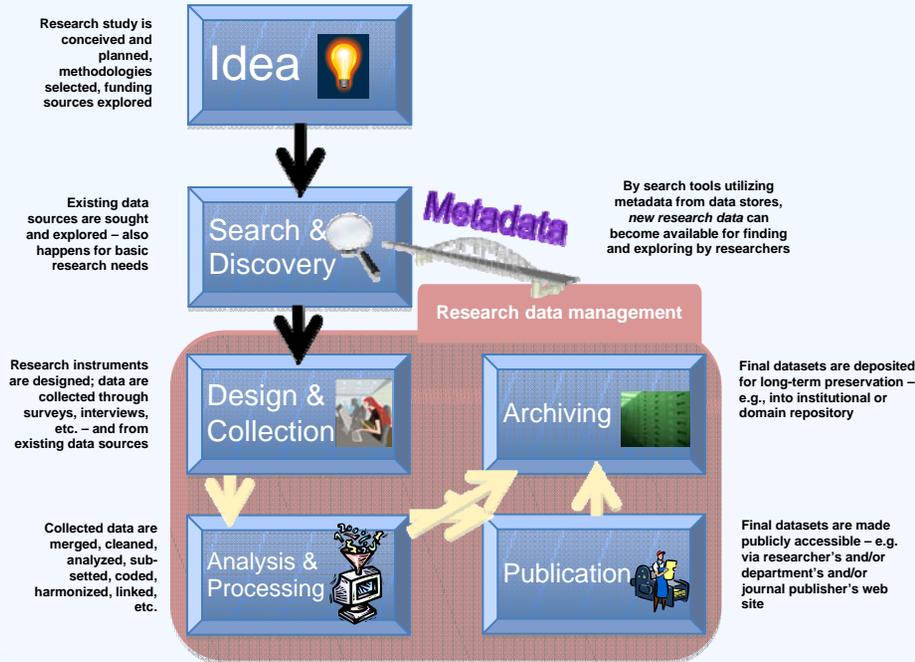


The Lifecycle of Social Science Research Data:

Improved Discovery through better Metadata *and* Search Tools

One research data management output is metadata preparation and its exposure to external search tools

Tomorrow's precise, metadata-driven, research data-focused **Search & Discovery**



Metadata for social science research data:

```
<r:Citation>
<r:Title>Consumer Expenditure Survey, 2004: Diary Survey</r:Title>
<r:Creator>U.S. Dept. of Labor, Bureau of Labor Statistics</r:Creator>
<r:Publisher>Washington, DC: U.S. Dept. of Labor, Bureau of Labor Statistics, 2005</r:Publisher>
<r:Publisher>[DIST] Ann Arbor, MI: Inter-university Consortium for Political and Social Research, 2006</r:Publisher>
<r:PublicationDate>
<r:SimpleDate>2006-02
</r:SimpleDate>
</r:PublicationDate>
</r:Citation>
<r:abstract id="">
```

Some search and discovery tools (among others) that might provide data-focused searching:

- WolframAlpha computational knowledge engine
- OAlster Find the pearls
- Google public data explorer labs
- BASE Berkeley Academic Search Engine

earch data management

```
<xs:element name="Geography" type="GeographyType"/>
<xs:element name="StartDate" type="BaseDateType"/>
<xs:element name="EndDate" type="BaseDateType"/>
<xs:element name="DataCollectionFrequency" type="DataCollectionFrequencyType"/>
<xs:element name="SamplingProcedure" type="r:IdentifiedStructuredStringType"/>
```



| | | |
|------------------------------|---|--------------------------------------|
| William C. Block Director | Stefan Krame Research Data Management Librarian | Jeremy William Programmer/Analyst |
|------------------------------|---|--------------------------------------|



Cornell University
Cornell Institute for Social and Economic Research