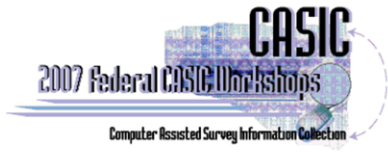
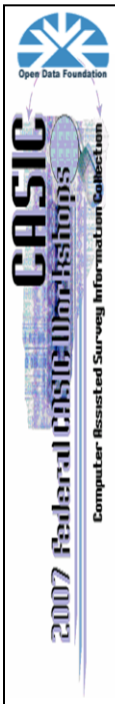


11th Annual Federal CASIC Workshops
Washington, DC, March 6 - 8, 2007
Session WP4

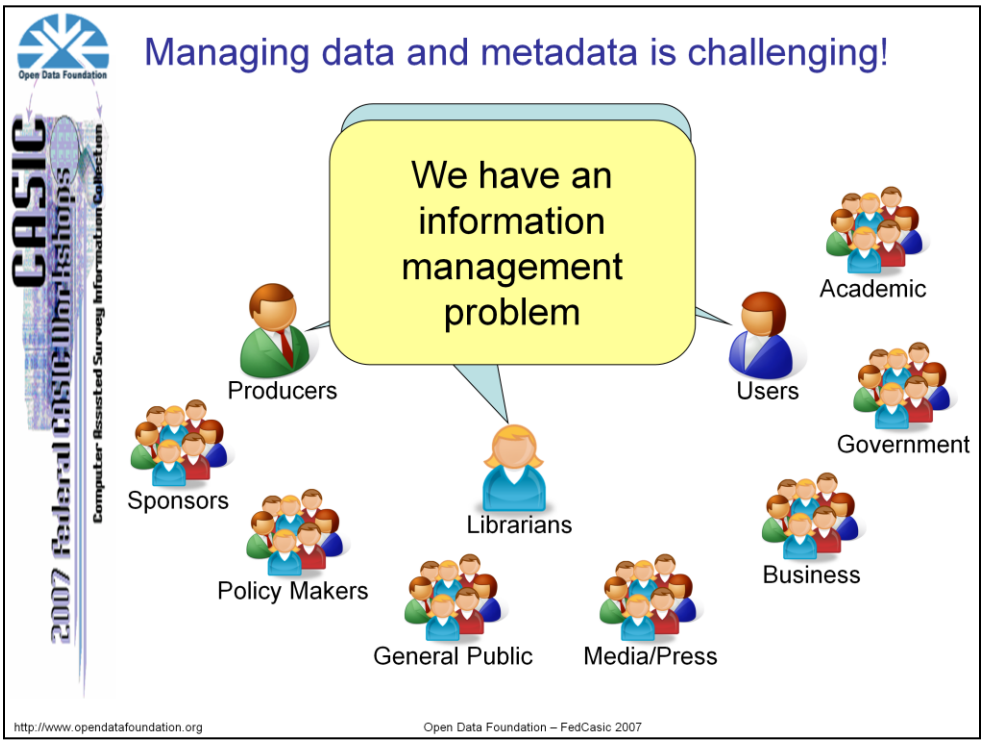
Metadata challenges and solutions for socio-economic data

Pascal Heus
Open Data Foundation
pheus@opendatafoundation.org
http://www.opendatafoundation.org



Outline

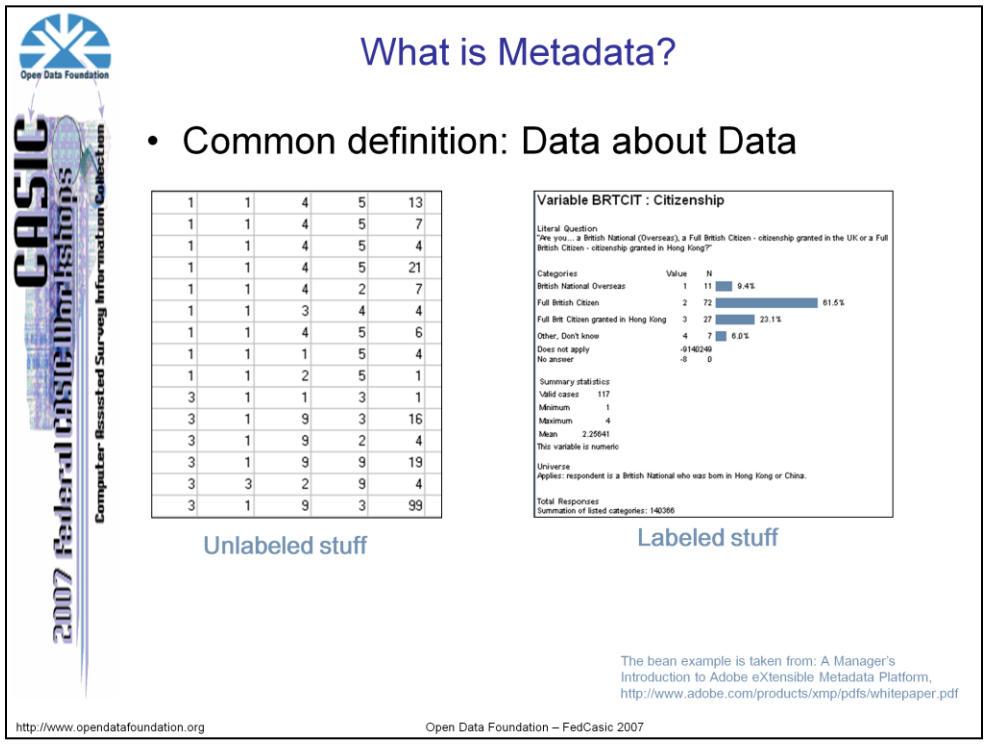
- Needs and challenges in statistical data and metadata management
- Metadata and XML solutions
- Selecting specifications
- Need for tools
- Open Data Foundation
- Conclusions / Q&A



Many actors & communities with different needs and perspectives

- Users: want open access to high quality and well documented data. Need discovery tools.
  - Public sector, private sector, academics
- Producers: prepare the data and need to comply with privacy laws
- Data Archives: need to interface with both communities
- Policy Makers: need data to measure results and impact and to plan ahead
- Sponsors: want to support the most relevant data collection
- Public and Media: want access to simple, easy to understand statistics

Solving Information management issues is what ICT & XML are for



-Provide descriptive information about of an object or concept

-Properties, characteristics (in XML: elements and attributes)

-It does not alter the content or nature of the object

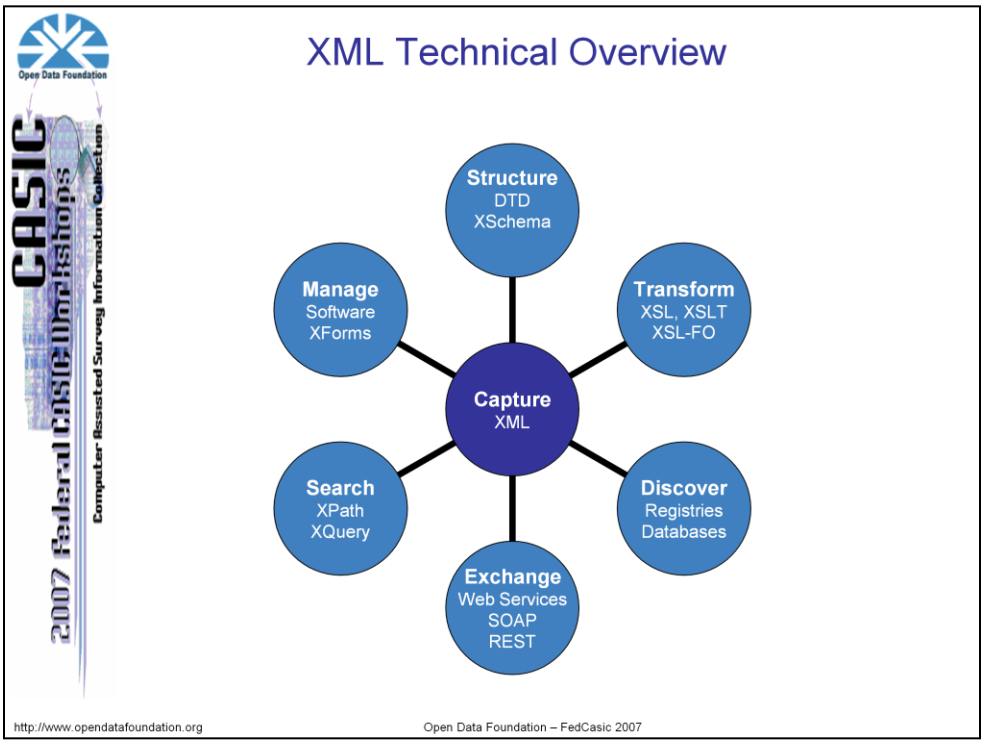
-It can be carried around without having to share the underlying object: catalogs, cars, libraries, etc.

-It is usually public domain (important for sensitive data)

XML to the rescue!

- XML is driving today's web service oriented architecture of the Internet and Intranets
- Using XML, we can capture, structure, transform, discover, exchange, query, edit and secure metadata and data
- XML is platform & language independent and can be used by everyone
- XML is both machine and human readable
- XML is non-proprietary, public domain and many open tools exist
- Domain specific standards are available!

- Technologies
  - Capture: XML
  - Structure: XSchema
  - Transform: XSL, XSLT, XSL-FO
  - Discover: Registries
  - Exchange: SOAP, REST, etc.
  - Query: XPath, XQuery
  - Edit: XForms
  - Secure: WS-Security (OASIS), etc.



XML

•Is e**X**tensible and is concerned with capturing information (unlike HTML who is not extensible a and focuses on representation)

•It's a **M**arkup system

•Is a **L**anguage with a syntax and grammar

XML is also a complete set of technologies for managing information/knowledge:

•Capture: metadata and/or data can be expressed using the XML language.

•Structure: Document Type Definition (DTD) and XSchema are use to validate an XML document by defining namespaces, elements, rules.

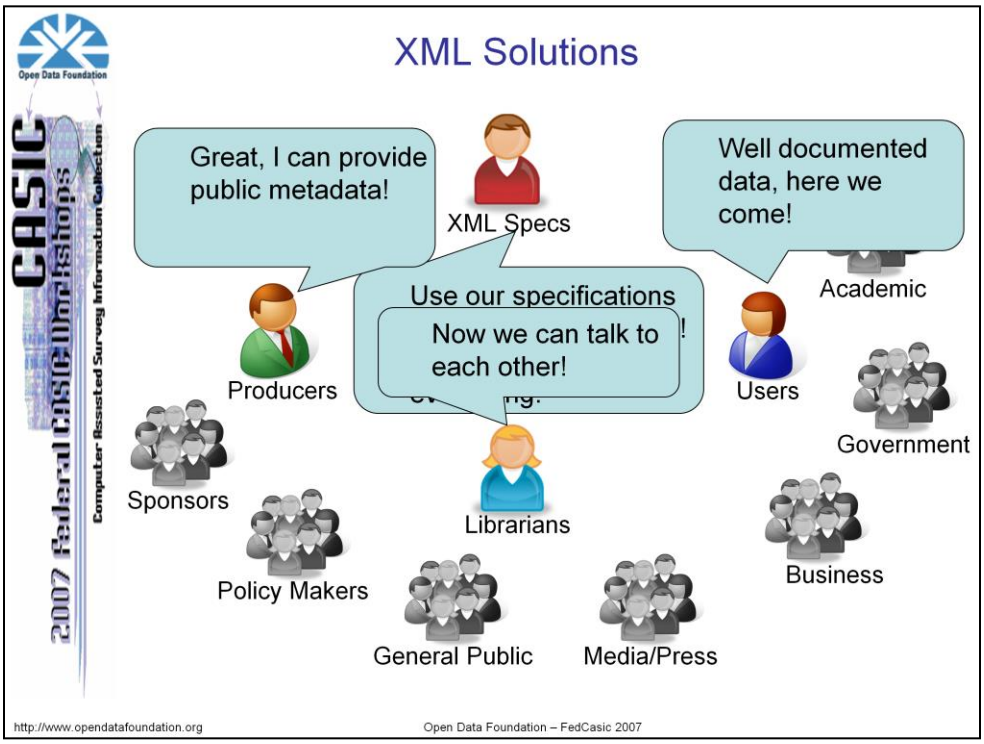
•Transform: XML separates the metadata storage from its presentation. XML documents can be transformed into something else, like HTML, PDF, XML, other) through the use of the

•Discover: using registries and/or native or relational databases

•Exchange: 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)

•Search: Very much like a database system, XML documents can be searched and queried through the use of XPath. There is no need to create or maintain tables, indexes or define relationships!

•Manage: Specialized software and can be used to create and edit XML documents. The XForms specification can also be used

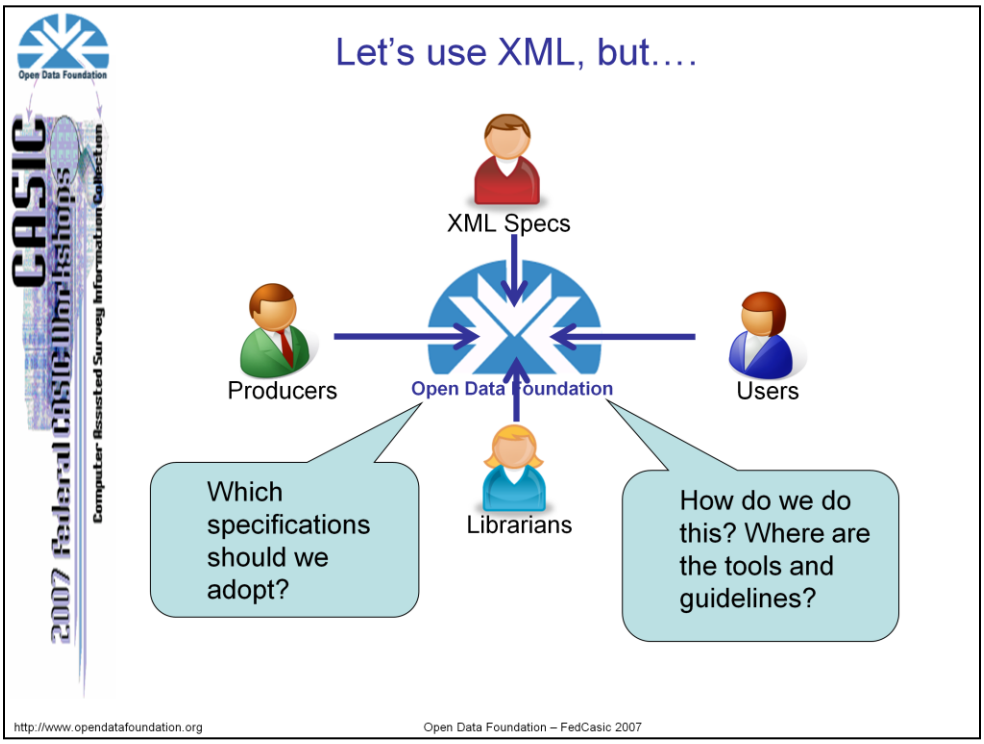


→A new actor: the specification/standard settings agencies, consortiums, alliances, etc.

→Use XML specifications will solve your problems

→User, Producers and Librarians have many reasons to cheer

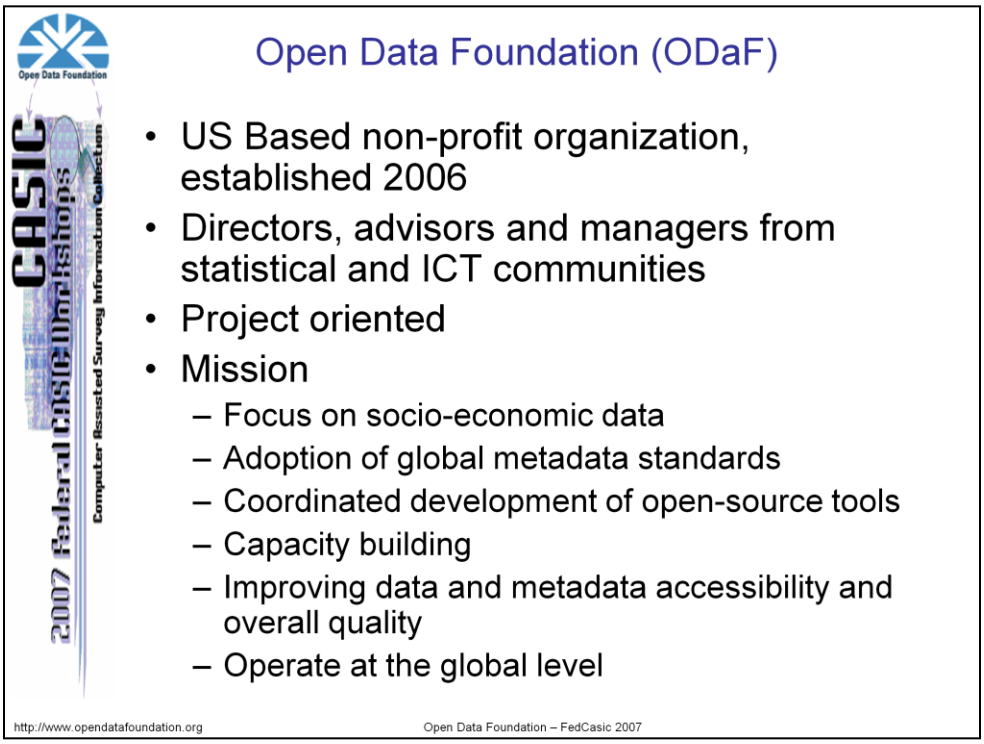
→But….



→ Perfect, let's use XML But…

- Which XML specification should we adopt?
- Where are the tools? How do we do this?

→ The Open Data Foundation has been established to answered these issues and needs



GLOBAL:

-Same issues present in all countries/agencies

-XML solutions are global solutions

-Data without borders: Global understanding of socio-economic issues requires global data (population/economic growths)

Directors:

Ernie Boyko - President of the International Association for Social Science Information Service and Technology (IASSIST)

Rune Gloersen - Head of Information Technology, Statistics Norway

Robert Glushko, PhD - Member of the OASIS Board of Directors, and the founder and leader of Berkeley's Center for Document Engineering

Julia Lane - Senior Vice President Director, Economics, Labor and Population, National Opinion Research Center (NORC) / University of Chicago

Advisors:

Sandra Cannon - Board of Governors of the Federal Reserve System

Gilles Collette - Visual Communications, Pan-American Health Organization

Daniel Gillman - US Bureau of Labor Statistics

Eduardo Gutentag - Member of the the OASIS Board of Directors

Paul Johanis - Statistics Canada

Graeme Oakley - Australian Bureau of Statistics

Ken Miller - UK Data Archive / Economic and Social Data Service

Juraj Riecan - United Nations Economic Commission for Europe (UNECE)

Gerard Salou - European Central Bank

Professor Bo Sundgren, Ph.D - Statistics Sweden

Wendy Thomas - Minnesota Population Center, University of Minnesota

Mary Vardigan - Inter-University Consortium for Political and Social Research

Management Team:

Arofan Gregory - specialist in SGML and XML-based open standards in the areas of publishing, e-commer ce, and statistics. Recent work includes participation in ebXML and related initiatives, and acting as a technical expert for SDMX and DDI.

Pascal Heus - an experienced IT specialist with a focus in microdata management systems. He has worked with international agencies such as the World Bank and the International Household Survey Network, and with national statistical agencies in developing countries. He is also active in the DDI initiative.

Chris Nelson - a modeling specialist who was a significant contributor to the OMG's Common Warehouse Metamodel, he has also worked for many years with GESMES (a statistical standard in EDIFACT syntax) and as a technical expert in the SDMX initiative.

Jostein Ryssevik - active within the DDI community, he was a key player in the development and sucess of Nesstar, the pre-eminent DDI-based toolkit.

Selecting XML specifications

- A single specification is not enough!
  - XML specifications commonly focus on a specific area of knowledge and/or set of functionalities
  - Cannot answer the needs of all actors
- XML mappings between specifications are possible
  - Information can be converted from one domain to another and be carried across communities
- Which ones should we use?
  - Fit for purpose
  - Widely accepted and supported
  - Can be mapped to a cross-domain family

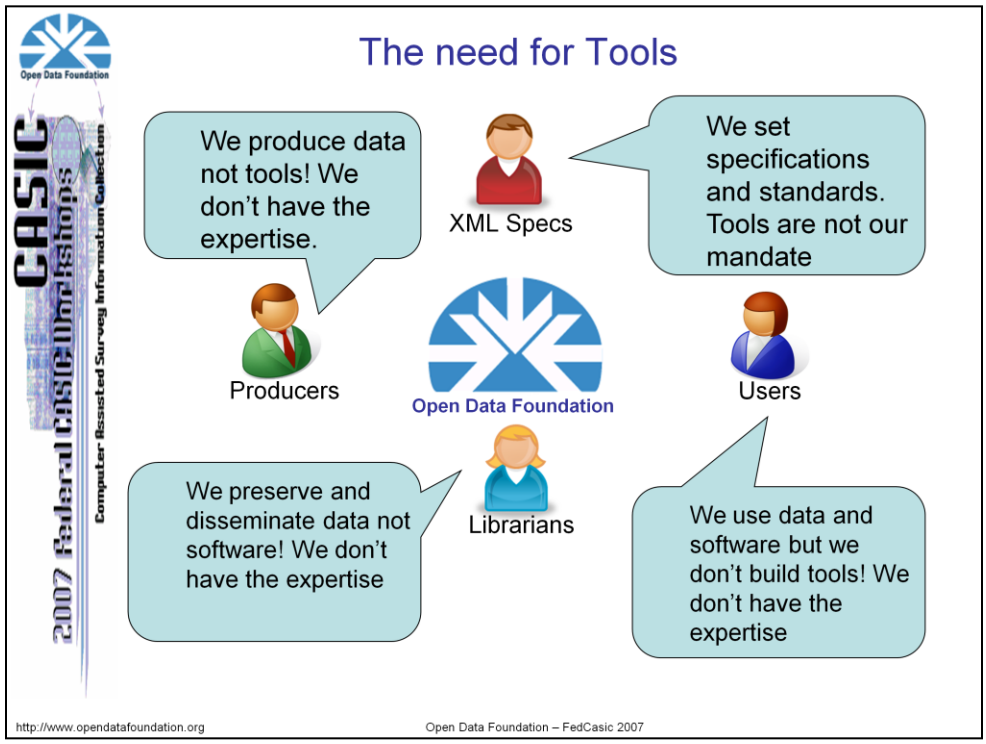
•Mappings: remember that , XML is easy to convert to another XML, it's build in the technology

A suggested set for socio-economic data

- Statistical Data and Metadata Exchange (SDMX)
  - Macrodata, time series, indicators, registries
  - http://www.sdmx.org
- Data Documentation Initiative (DDI)
  - Microdata (surveys, studies)
  - http://www.ddialliance.org
- ISO 11179
  - Semantic modeling, concepts, registries
  - http://metadata-standards.org/11179/
- ISO 19115
  - Geography
  - http://www.isotc211.org/
- Dublin Core
  - Resources (documentation, images, multimedia)
  - http://www.dublincore.org

This is a set of specifications for socio-economic data

When it comes to implementation, these are complemented with commonly used ICT specifications such as the XML family of recommendations, SOAP, OASIS WS-* security specifications, SVG, etc.



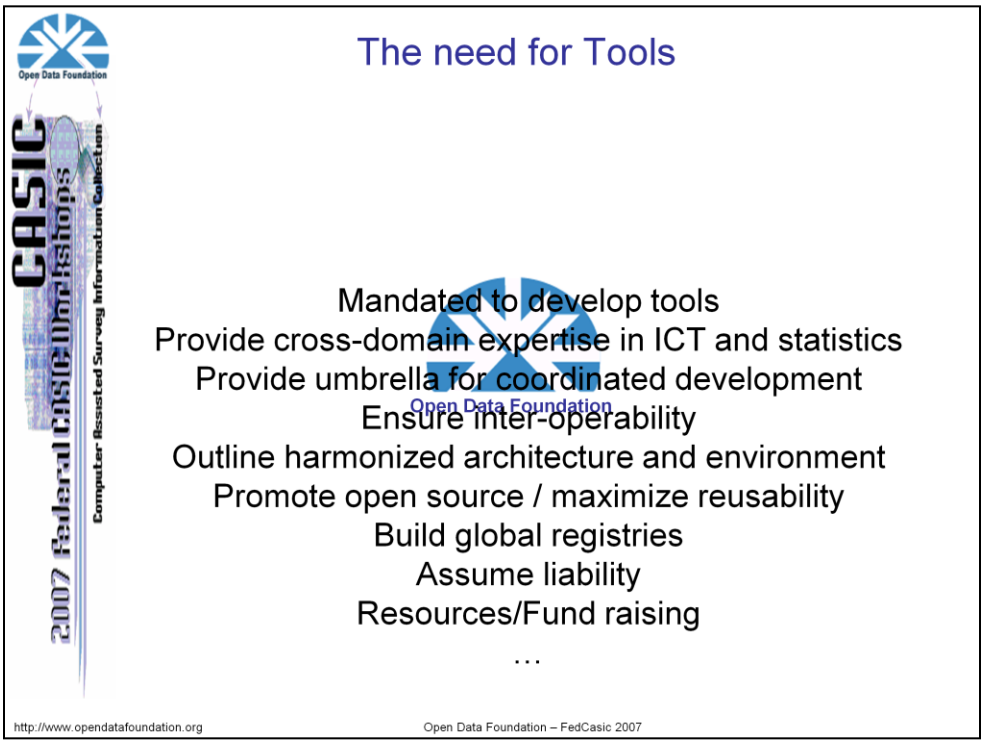
Software application and guidelines are crucial for the adoption of XML specifications

But very few organizations are developing such tools:

- •Lack of mandate: most of the agencies are not in the business of developing software
- •Lack of expertise: even if the would want to, they seldom have the ICT capacity to do so
- •Lack of coordination: agencies are often locked into their own world and are not particularly interested in the big picture. Someone must be there to coordinate efforts and ensure compatibility
- •Lack of funding: since the mandate is not there, the money rarely follows. We need a way to raise awareness and funding for tool development
- •Liability issues: agencies do not want to be held responsible

→ Open Data Foundation

•Coordinated development of open source tools in an harmonized framework



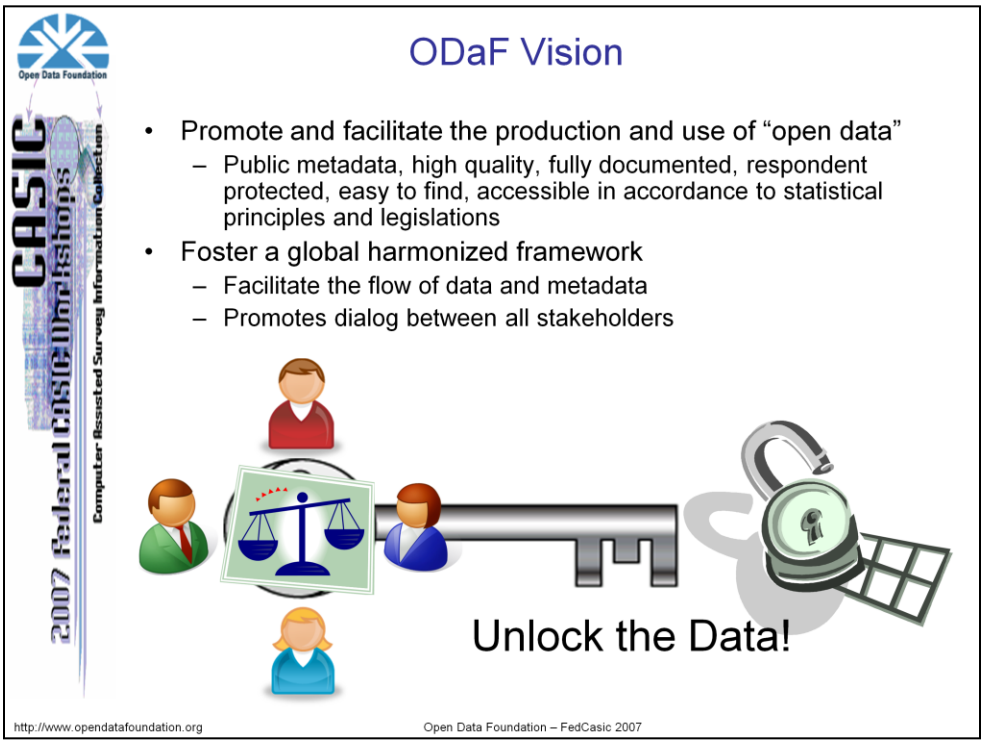
Software application and guidelines are crucial for the adoption of XML specifications

But very few organizations are developing such tools:

- •Lack of mandate: most of the agencies are not in the business of developing software
- •Lack of expertise: even if the would want to, they seldom have the ICT capacity to do so
- •Lack of coordination: agencies are often locked into their own world and are not particularly interested in the big picture. Someone must be there to coordinate efforts and ensure compatibility
- •Lack of funding: since the mandate is not there, the money rarely follows. We need a way to raise awareness and funding for tool development
- •Liability issues: agencies do not want to be held responsible

→ Open Data Foundation

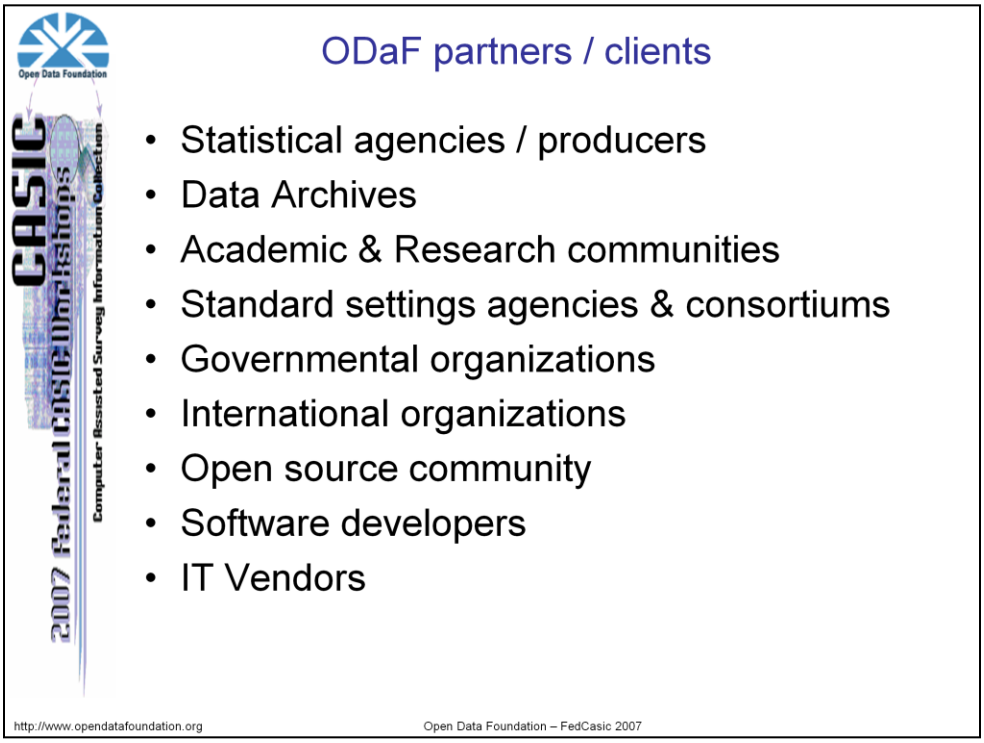
•Coordinated development of open source tools in an harmonized framework



•The harmonized framework is the key to unlock the data

Some ODaF Projects & Ideas

- Guidelines for an harmonized architecture and development environment
- Develop tools for agencies
- XML mappings
- Facility to host development of open source projects (GForge)
- Provide hosting services for agencies
- Implement registries
- Produce training and reference material
- Technical support & capacity building
- …



A few examples:

•DDI Alliance

•US Census

•Interuniversity Consortium for Political and Social Research (ICPSR)

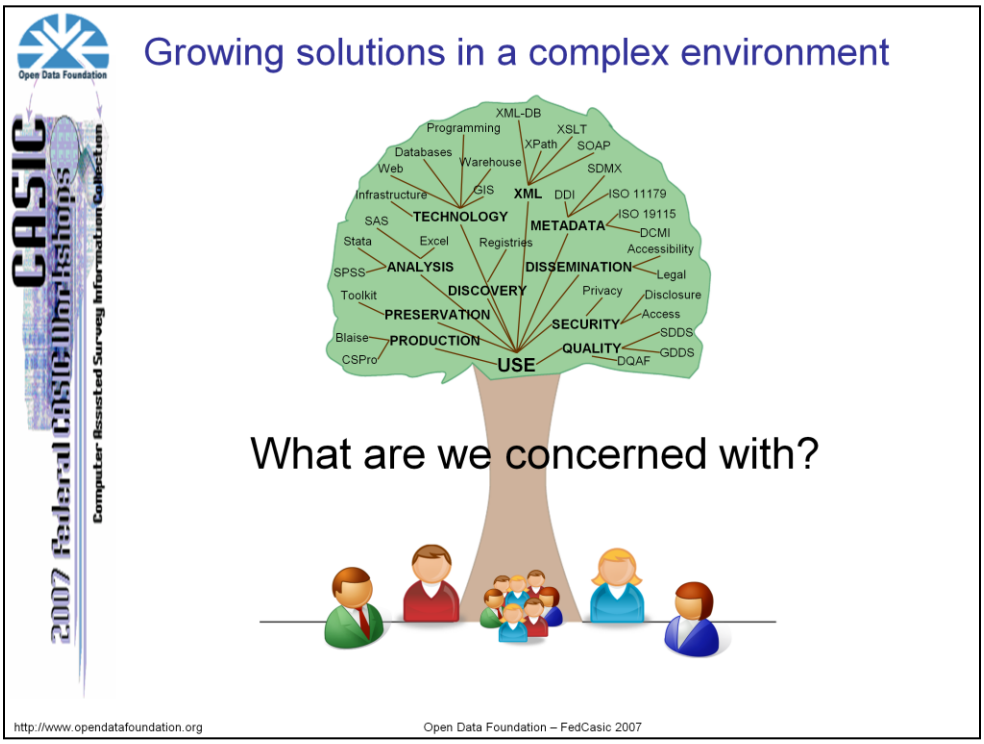
•National Opinion Research Center (NORC)

•UNESCO Institute for Statistics

•International Household Survey Network

•Food and Agriculture Organization (FAO)

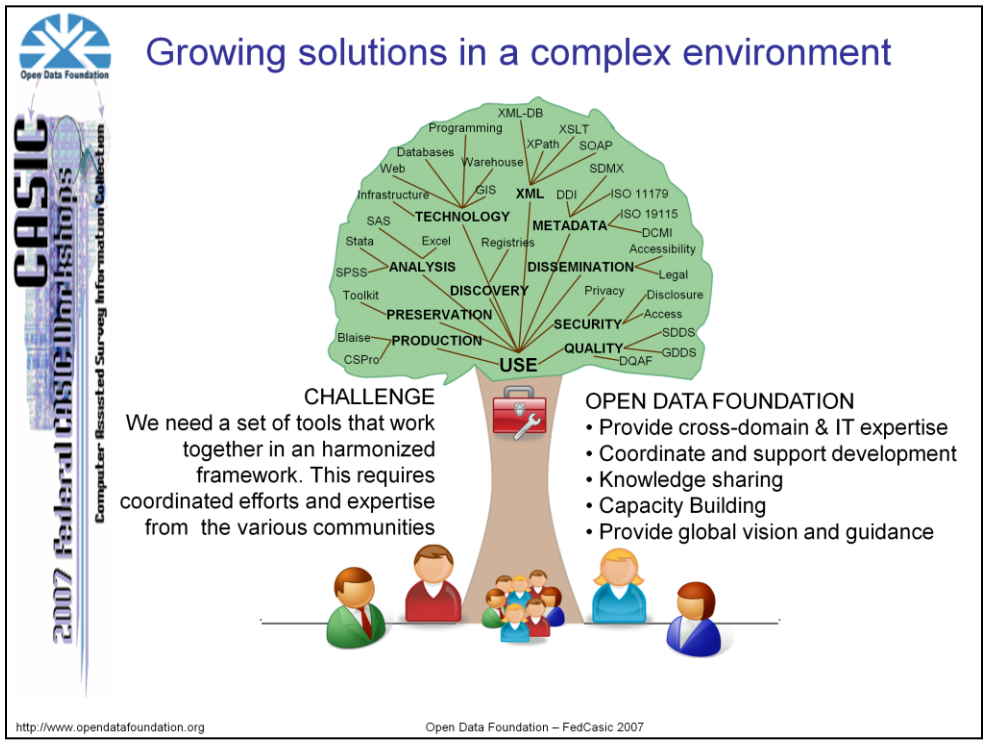
•UN Economic Commission for Europe (UNECE)



-We are looking at the many dimensions of the socio-economic landscape

-Rooted in several communities

-Need to grow solutions in a very complex environment (through communities)



We need a harmonized toolbox

ODaF's role:

-Provide cross-domain expertise

-Coordinate and support development of open tools

-Share knowledge

-Capacity Building

-Provide a global vision

ODaF Challenges

- The technology is available today
- The right people are available today
- The need and the will are there
- The real challenges are:
  - Awareness / Understanding of technology
  - Change management
  - Content management
  - Coordination & Guidance
  - Focused resources and funding
  - Institutional commitment
- Learn for the past for a better future
- It's not about data, it's about people

http://www.opendatafoundation.org Open Data Foundation – FedCasic 2007

Change management

•This is the biggest challenge

•Need to overcome resistance to change and take the first steps

Change management

•Collecting and compiling the right information

•Metadata quality control

Awareness & Understanding

•Need to be aware that solutions exists

•Need to understand what the technology can do

•Promote ODaF and partners

Coordination & Guidance

•Need for the right expertise

•Need for training, best practices, champions

Focused resources and fund raising

- Need to commit resources to achieve this
- It won't be free but certainly worth the return on investment!
- Investment is a small percentage of what is invested in the data production efforts

Institutional commitment

- Success cannot be achieved individual efforts only, institutions and people need to come together
- Successful projects require upper management support

Rapid results are possible!

Learn from the past for a better future:

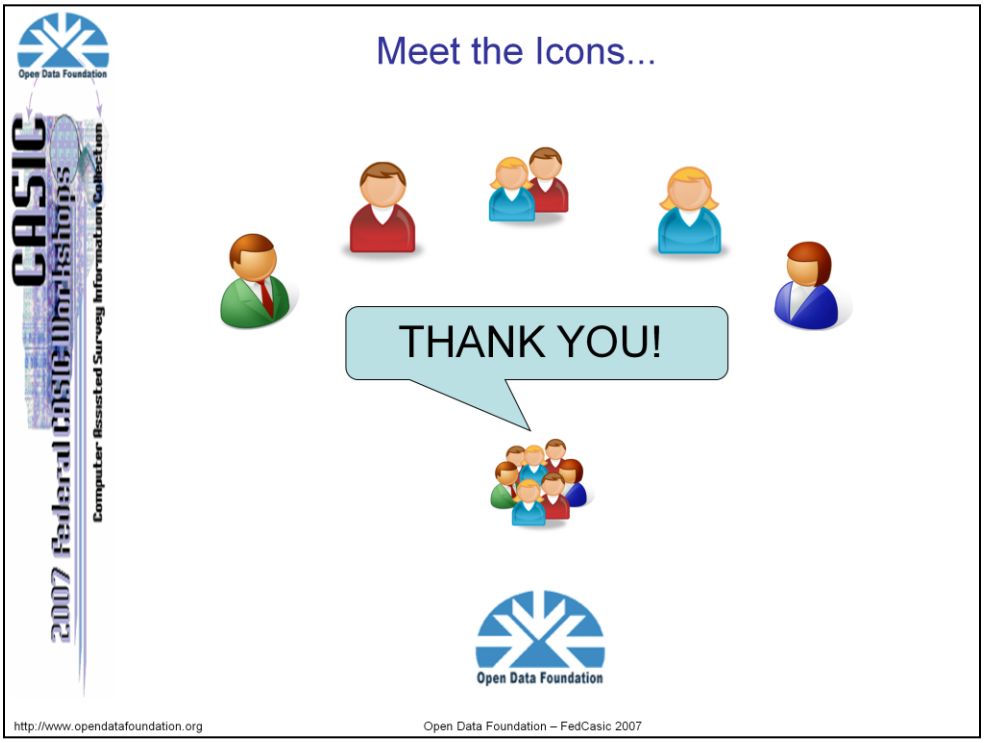
- We cannot change what has been produced & done so far but we can decide for a different future today
- We've done the best we can so far. Let's accept that we have not been perfect, transparency is vital.
- Integrating new tools and techniques in the data life-cycle will improve the overall quality and usefulness

Our data is about people

- We need to develop a sense of urgency, the sooner the better
- Policy makes need access to better data for better results
    - Evidence based policies
    - Results based framework
- This has serious impact on living conditions

Summary

- Managing data and metadata is challenging
  - Solutions exist to make it easier and provide better information to unlock the data
- Adopt a set of specifications that answer your requirements and can connect across domains
  - DDI, SDMX, ISO 11179, Dublin Core, ISO 19115
- Promote the use and development of open tools, do not work in isolation, get the appropriate expertise
  - Open Data Foundation


Open Data Foundation
CASIC
2007 Federal CASIC Workshops
Computer Assisted Survey Information Collection
Meet the Icons...
THANK YOU!
Open Data Foundation
http://www.opendatafoundation.org
Open Data Foundation – FedCasic 2007