SDMX and DDI: How Do They Fit Together in Practical Terms?

Arofan Gregory
The Open Data Foundation
European DDI User's Group 2011
Gothenburg, Sweden

Outline

- Background
- Characterizing the Standards
  - DDI
  - SDMX
  - Similarities and Differences
  - Other Relevant Standards
- Implementation Approaches
  - DDI In, SDMX Out
  - SDMX-Centric
  - Standards Agnostic
- Future Possibilities: The SDMX-DDI Dialogue Proposal

Background

- This presentation intends to examine the different *architectural approaches* to implementations of SDMX and DDI together
  - While several organizations are mentioned, it is not a report on the status of prototypes or implementations
- This presentation does *not* intend to introduce DDI or SDMX to an unfamiliar audience
  - Familiarity with the standards is assumed

Background (2)

- When people think about using SDMX and DDI together, they make assumptions
  - Microdata (and tabulations) can be described using DDI
  - A transformation could be applied to produce SDMX to describe the aggregates/tables
  - There is a straight mapping from DDI to SDMX
- Interestingly, this conceptual model is not how the use of DDI and SDMX together is being approached in reality
  - The Devil is in the details! (Or is it “The Tomten is in the details” ?)

Background (3)

- People have been discussing the use of SDMX and DDI together for some time
- Now, we are at the stage where implementations are being investigated and prototyped
  - Not “if”, but “how”
- Most often, this is done in the context of the Generic Statistical Business Process Model (GSBPM), by data producers
  - The idea of “industrialized” statistical production
  - Strong emphasis on process management

Characterizing the Standards: DDI

- DDI Lifecycle can provide a very detailed set of metadata, covering:
  - The study or series of studies
  - Many aspects of data collection, including surveys and processing of microdata
  - The structure of data files, including hierarchical files and those with complex relationships
  - The lifecycle events and archiving of data files and their metadata
  - The tabulation and processing of data into tables (Ncubes)
    - Allows for a link between the microdata variables and the resulting aggregates

Characterizing the Standards: SDMX

- Describes the structure of aggregate/dimensional data ("structural metadata")
- Provides formats for the dimensional data
- Provides a model of data reporting/collection and dissemination
- Provides a way of describing the structures of arbitrary metadata sets ("reference metadata")
- Provides formats for the arbitrary metadata sets
- Provides a set of standard registry interfaces, providing a catalog of resources
- Provides guidelines for deploying standard web services for SDMX resources
- Provides a way of describing statistical processes

Differences

- DDI has much more detailed metadata at the level of the study, because it is intended to describe the full process of data production (the data lifecycle)
- DDI provides more complete descriptions of the processing of data
- SDMX provides more architectural components, to support reporting/collecting and exchange

Similarities: Design

- Both standards use a similar mechanism for structuring URN identifiers
- Both standards use a similar model for identifiable, versionable, and maintainable things
  - Both have a concept of an owning agency
  - There is a very similar set of rules about versioning and maintenance
- Both standards use “schemes” as packages for lists of like items
- Both standards are designed to support reuse, and have similar referencing models

Similarities: Specific Metadata Items

- Concept Schemes
- SDMX Codelists/DDI Codes and Categories
- Dimensional data structures (Ncubes/DSDs)
- Organization Schemes

- There is an effort as part of the SDMX-DDI Dialogue to produce a common vocabulary of terms, describing similarities and differences

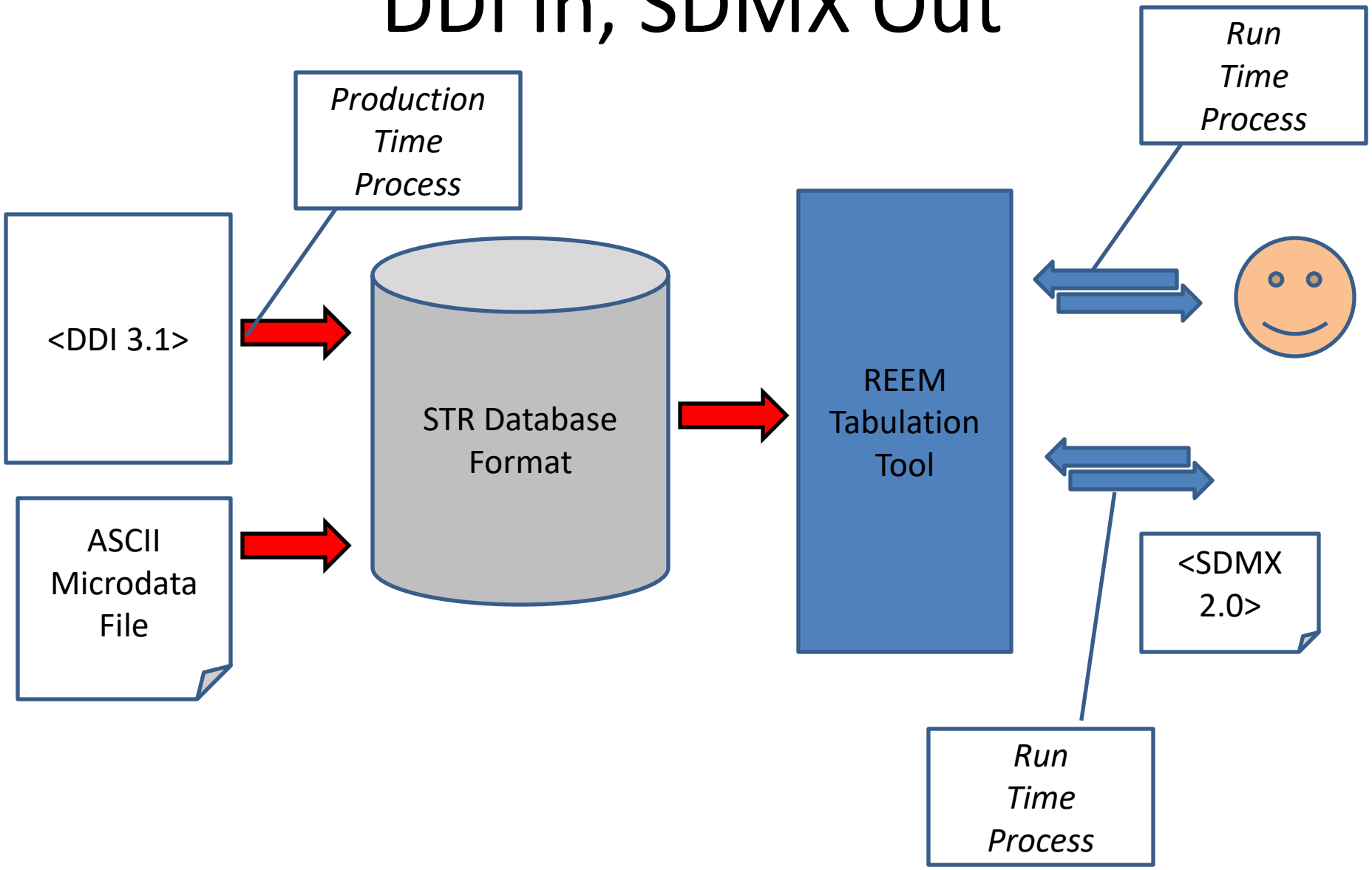
Other Relevant Standards

- Some things are not covered well by either SDMX or DDI, particularly classification management
  - The Neuchatel model is probably a better standard, but it has no standard XML representation
  - The older (and similar) CLASET model is also potentially useful, and does have an XML representation
- The GSBPM gives us a generic model for describing business processes, but to implement process management you will use other standards such as BPMN (specific process modelling) and BPEL (for executing processes)

Implementation Approach: DDI In, SDMX Out

- This is an approach used by the Australian Bureau of Statistics (ABS) in one part of their microdata access facility, REEM
- It is based on a set of software tools developed and sold by Space-Time Research ("SuperCross") to support tabulations from microdata

DDI In, SDMX Out

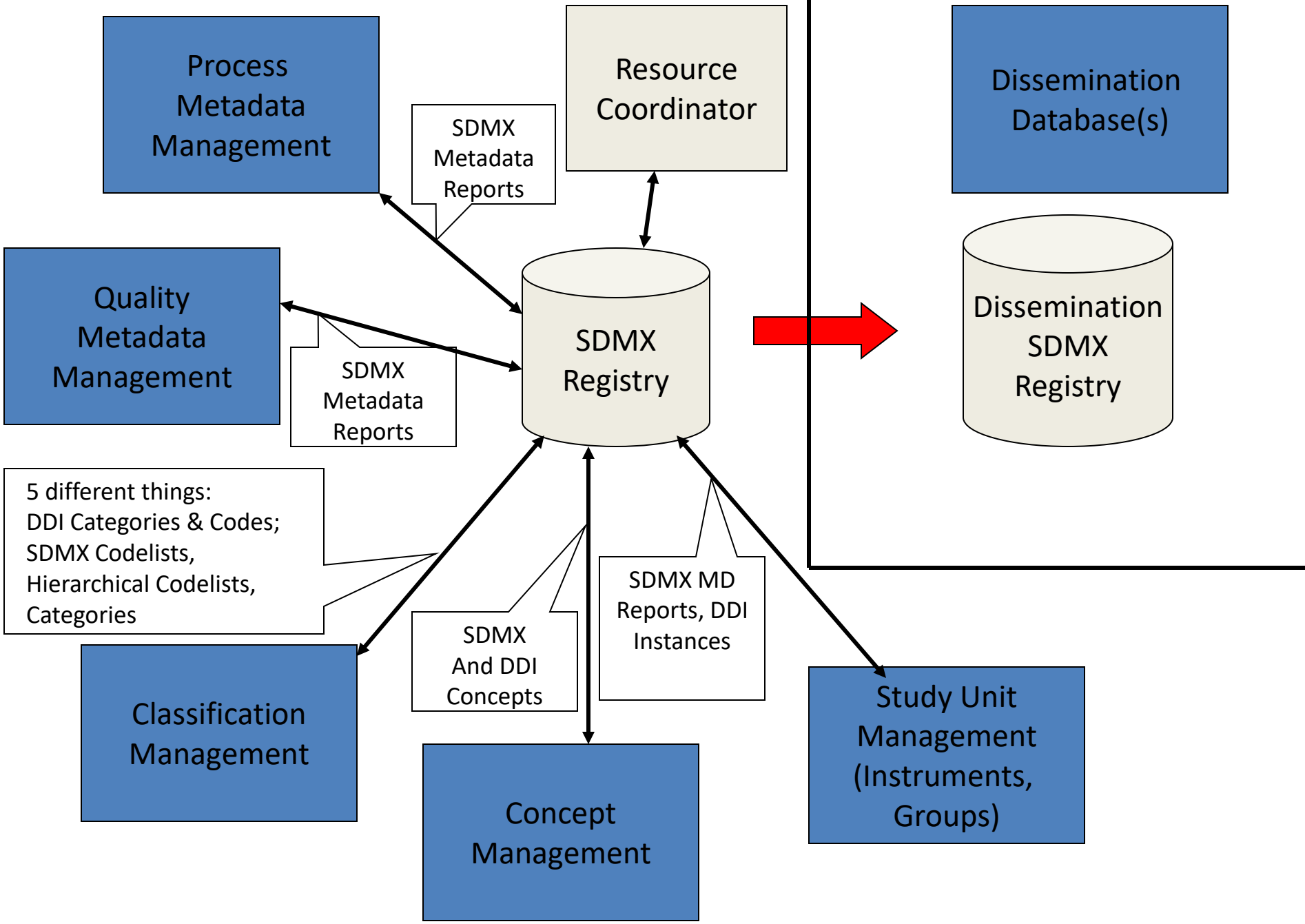


Considerations

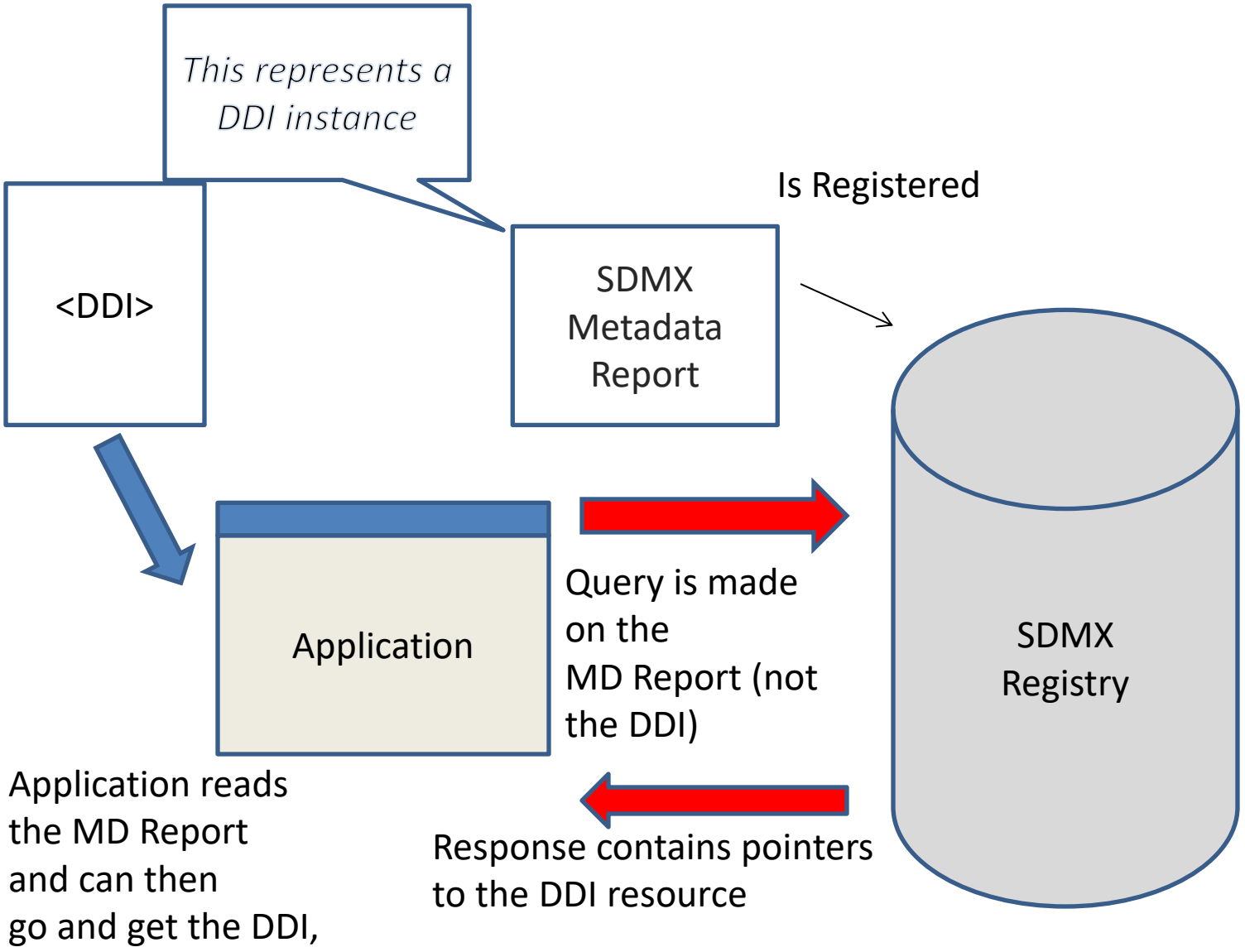
- This is a limited implementation, providing secure access to microdata in the form of user-defined tabulations
  - It is only a limited dissemination scenario
  - It relies on run-time confidentialization, which limits the data that can be made available (only data and tabulations for which robust automated confidentialisation can be assured)
- The internal formats are proprietary, and lack some of the richness of the DDI-L model
  - We also identified some bugs in DDI 3.1
- Not sufficient for users who wish to perform statistical analysis of the microdata rather than produce tabulations
  - That need will be met by another part of the overall REEM solution in future
- There is no direct mapping from DDI to SDMX

Implementation Approach: SDMX-Centric

- This approach came out of discussions within INSEE, as they considered designs for the new metadata repository they are developing
- Similar approaches have been considered by other organizations
- This relies heavily on the use of the SDMX architectural components and model, especially the SDMX Registry
- There is an idea of GSBPM-based process management, but no process-management tool


Process Metadata Management
SDMX Metadata Reports
Resource Coordinator
Dissemination Database(s)
Quality Metadata Management
SDMX Metadata Reports
SDMX Registry
Dissemination SDMX Registry
5 different things:
DDI Categories & Codes;
SDMX Codelists,
Hierarchical Codelists,
Categories
SDMX And DDI Concepts
SDMX MD Reports, DDI Instances
Classification Management
Concept Management
Study Unit Management (Instruments, Groups)

MSD for DDI

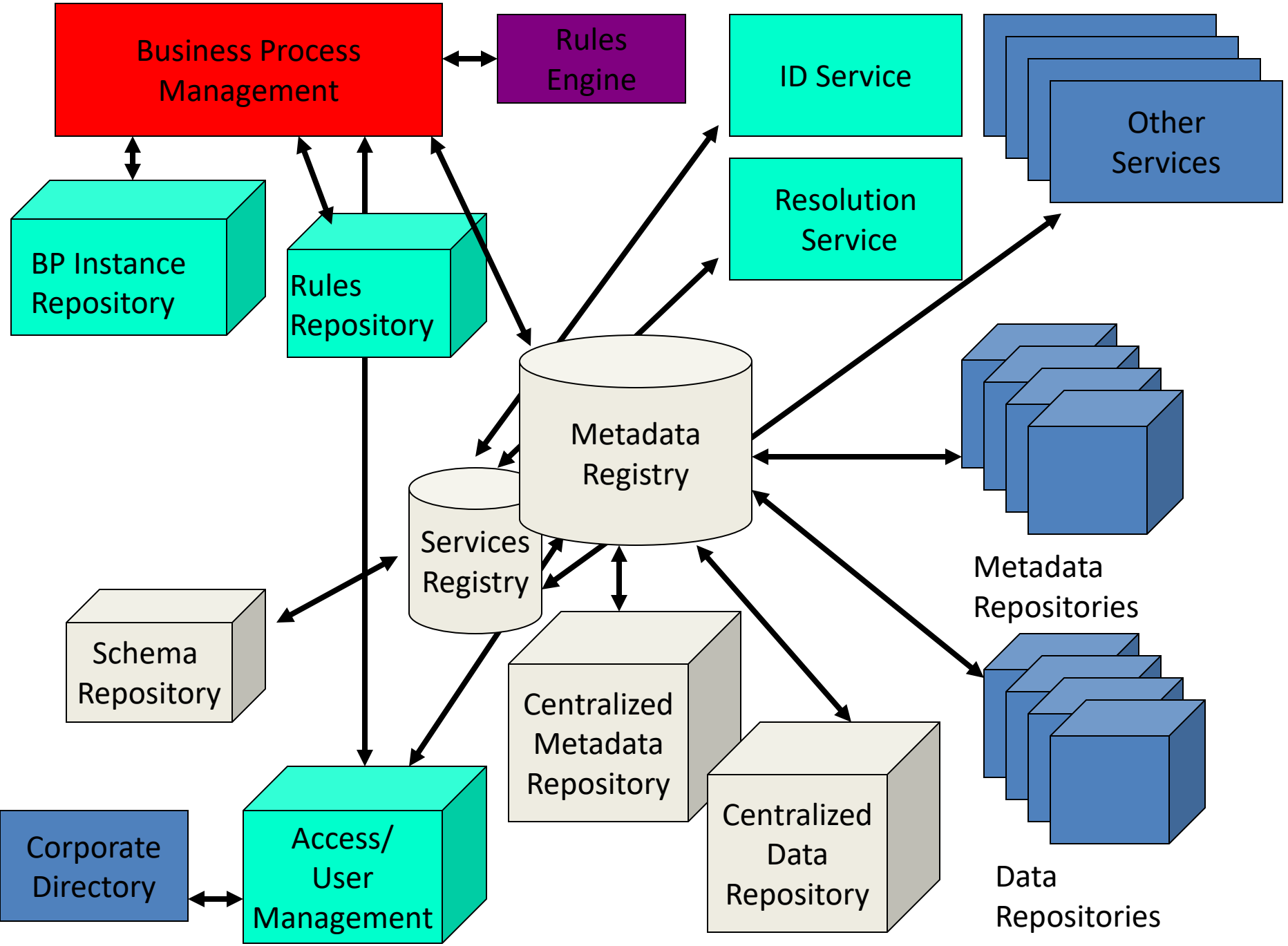

This represents a DDI instance
<DDI>
SDMX Metadata Report
Is Registered
Application
Query is made on the MD Report (not the DDI)
SDMX Registry
Application reads the MD Report and can then go and get the DDI,
Response contains pointers to the DDI resource

Considerations

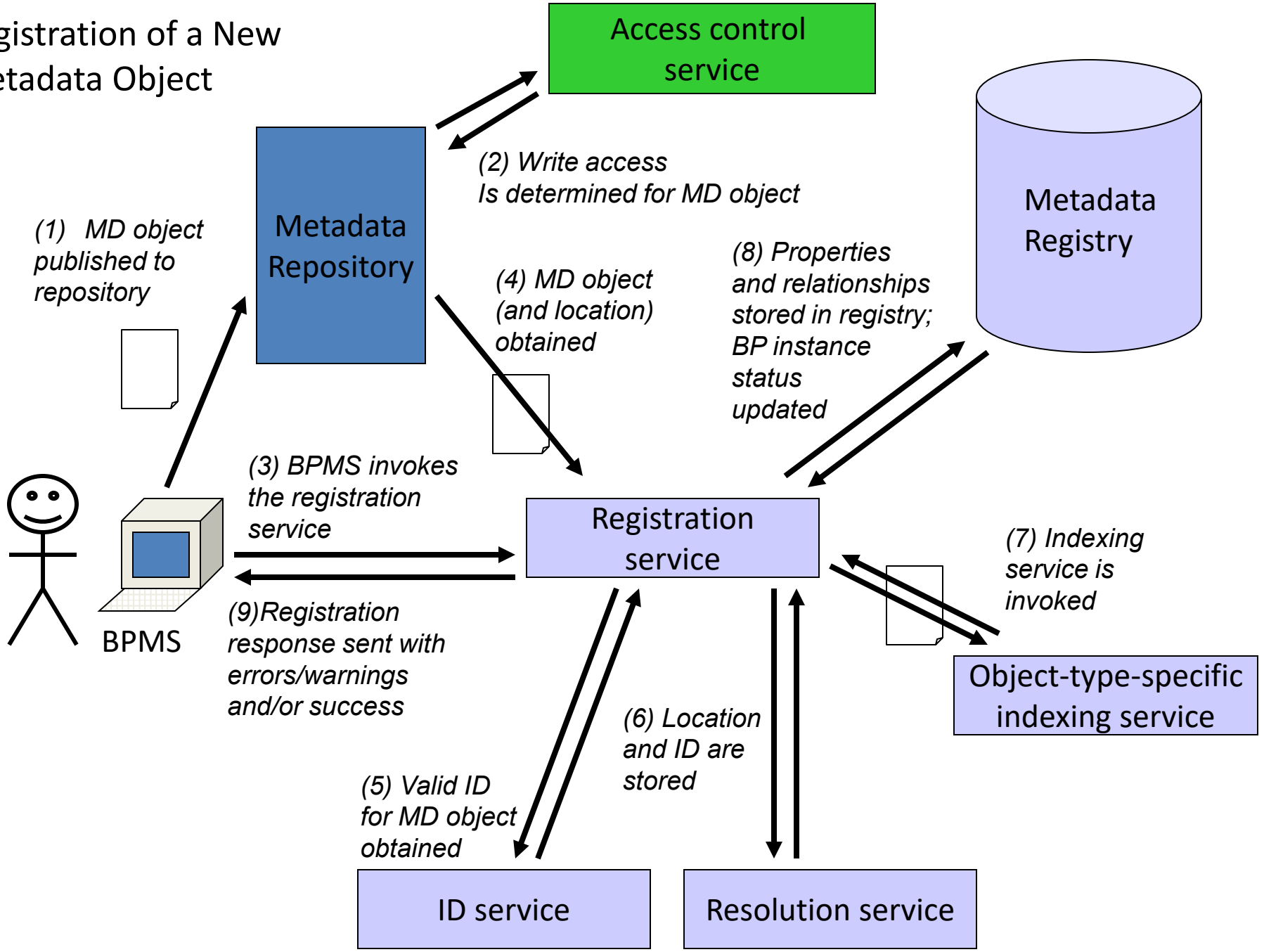
- The SDMX Registry is available as a free tool, reducing the amount of development needed to deploy such a system
  - Other SDMX tools are also available for free
- Applications are coded against specific versions of the standards, coming with fairly high maintenance costs if future versions need to be supported
- Access to non-SDMX resources (DDI) involves a level of indirection
  - Retrieval is a two-step process: first get the “placeholder” SDMX Metadata Report, process it, and then retrieve the non-SDMX resource (DDI)
- The GSBPM was described as a set of SDMX Processes, and these are held in the registry to help organize and manage the statistical production process

Implementation Approach: Standards Agnostic

- This approach is currently being prototyped by the ABS, as part of a major re-development of their IT infrastructure to “industrialize” their production processes
- It is a registry-based, distributed model, but it does not rely on the SDMX Registry, but on a standards-agnostic registry
- It is also based on the GSBPM, and on the emergent sibling to it, the Generic Statistical Information Model (GSIM)
- There is a major component of process management and automation


Business Process Management
Rules Engine
ID Service
Other Services
BP Instance Repository
Rules Repository
Resolution Service
Metadata Registry
Services Registry
Metadata Repositories
Schema Repository
Centralized Metadata Repository
Centralized Data Repository
Data Repositories
Corporate Directory
Access/ User Management

Registration of a New Metadata Object



Standards Agnosticism

- The term “standards-agnostic” means that the standards themselves are represented as metadata objects within the registry
  - Each version of each relevant standard is described as either a read-only or a sufficient read-write format for any type of object
  - Every metadata object describes which versions of which standards are supported
  - Transformation services between standards and versions are also registered resources
  - Introducing new standards or new versions of new standards has a minimal impact on existing applications
  - Some “standards” could be agreed organization-wide standards, not necessarily public standards such as DDI and SDMX

Considerations

- There is a huge emphasis on process automation and management
  - All functionality is exposed as web services – this is an “SOA” architecture which works well with existing process management tools
- The cost of developing and deploying the new infrastructure will be very high
  - Migration from legacy systems will be challenging
  - Organizational change issues will need to be overcome
- The value of deploying such a system will be immense
  - Flexibility and speed will be greatly increased for statistical production
  - Management of the statistical production process will be easier and more effective
  - Consistency and quality of the data products will be enhanced

Future Possibilities: The SDMX-DDI Dialogue Proposal

- There has been a set of informal meetings between members (and prospective members) of the SDMX community and the DDI community, looking for ways in which the standards can be used together effectively
  - The first meeting was held at EDDI 2010
  - There have been several other meetings since
- One proposal is now being discussed which outlines an approach to using SDMX and DDI interchangeably

A Simple Fact

- Its not about which flavor of XML you use – XML doesn't really matter
- It's about the *data* and the *metadata*!

The Challenge

- If I want to use DDI to describe my data, and you want to use SDMX, how can we ensure that we are getting the same data and metadata?

The Proposed Approach

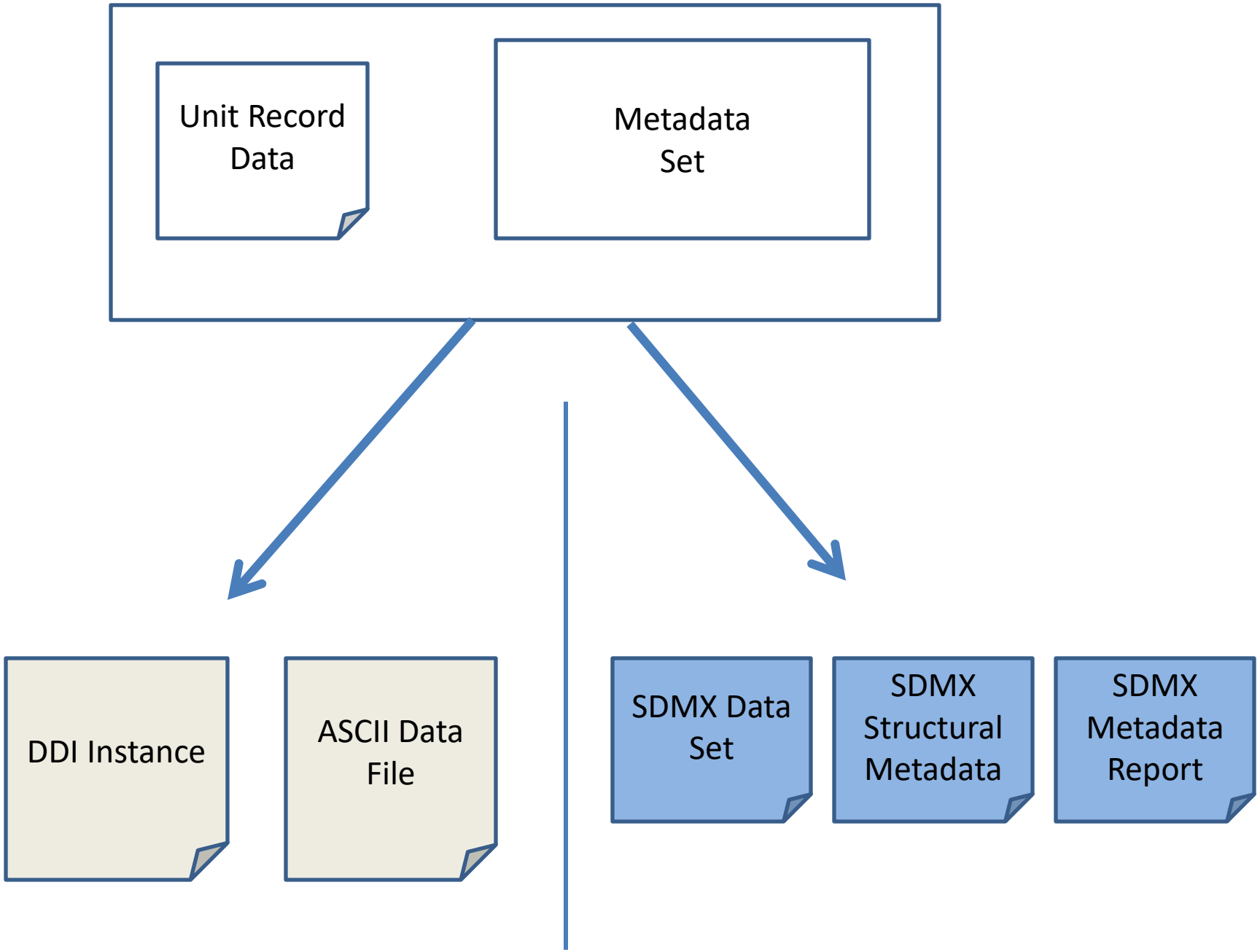
- The SDMX-DDI Dialogue has been defining a set of relevant business cases where the two standards could be used together
- One of these business cases involves retrieving unit record data from a register
- A model of the full set of useful data and metadata has been identified
  - The metadata is a subset of the DDI elements, which could be expressed in DDI as a “DDI Profile”

The Proposed Approach (2)

- The full set of information includes:
  - The unit record data
  - Structural information about the variables and representations
  - Additional information about how the data has been generated/collected/processed
- In DDI, this set of information can be expressed as a DDI instance and a data file
  - Both the structural and processing metadata can be expressed as a single DDI instance

The Proposed Approach

- In SDMX, we have three XML files:
  - A file holding the data, expressed as dimensional microdata
    - The unit identifier is a dimension
    - The variable identifier is a dimension
    - There are dimensions related to time
  - A reference metadata report will all other metadata describing the process/collection/generation of the administrative data
  - A file describing the concepts, data structure, and codelists (“structural metadata”) for the data, and also the structure of the metadata report


Unit Record Data
Metadata Set
DDI Instance
ASCII Data File
SDMX Data Set
SDMX Structural Metadata
SDMX Metadata Report

Results

- If I am using SDMX, but I am sent DDI, a simple transformation will give me the same payload of data and metadata
- Vice-versa for SDMX users
- There are some conventions which will need to be established regarding identifiers and the way the unit record files are structured
- There will need to be agreed models for each business case

An SDMX File?

```
<DDIInstance>
    <StudyUnit>
            <Agency>mpc.umn.us</Agency>
            <ID>23576</ID>
            <Version>1.0</Version>
            <Abstract>This is a description of the data sourced from the US Employment Service Register of Working Ar
            <UniverseReference>
                    <ID>Univ23576</ID>
            </UniverseReference>
            <SeriesStatement>This data collection is part of the ongoing data collection for the OECD's labor amd emp
            <Purpose>This data is collected under the agreement between the OECD and all member countries...</Purpose
            <Coverage>
                    <Topical>Labor and employment statistics</Topical>
                    <Spatial>The United States of America</Spatial>
                    <Temporal>2010</Temporal>
            </Coverage>
            <AnalysisUnit>Individual</AnalysisUnit>
            <KindOfData>Administrative data</KindOfData>
            <ConceptualComponents>
                    <UniverseScheme>
                            <Agency>OECD</Agency>
                            <ID>Univ23685</ID>
                            <Version>1.0</Version>
                            <Universe>
                                    <ID>Univ23576</ID>
                                    <Description>the working population of the United States.</Description>
                            </Universe>
                    </UniverseScheme>
            </ConceptualComponents>

            ....

    </StudyUnit>
<DDIInstance>
```

Questions?