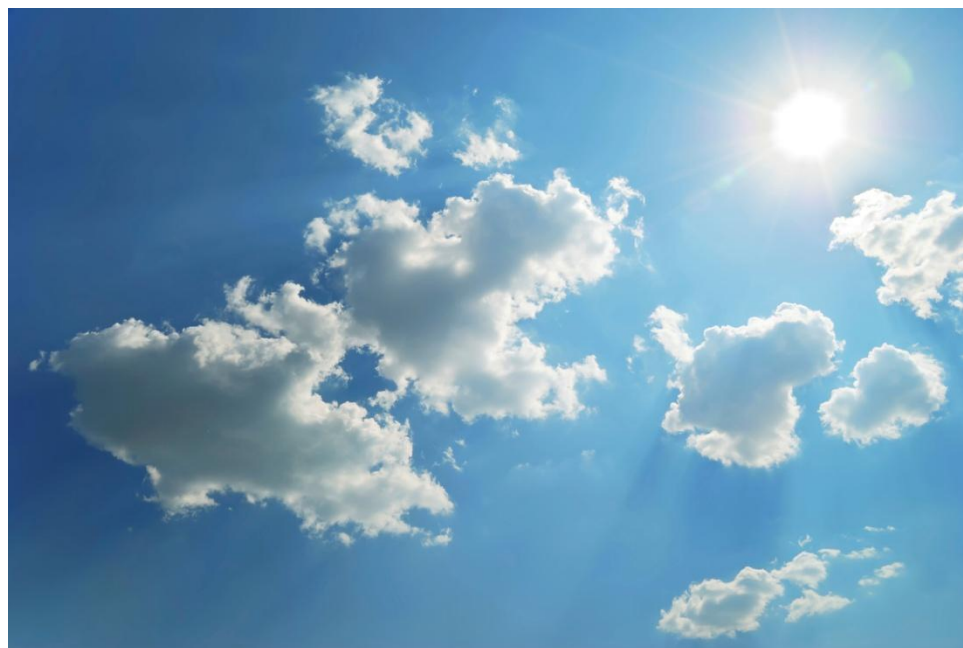


METADATA FOR THE LONGITUDINAL DATA LIFE CYCLE



By Larry Hoyle, Fortunato Castillo, Benjamin Clark, Neeraj Kashyap, Denise Perpich, Joachim Wackerow, and Knut Wenzig

03/31/2011

DDI Working Paper Series – Longitudinal Best Practice, No. 3

This paper is part of a series that focuses on DDI usage and how the metadata specification should be applied in a variety of settings by a variety of organizations and individuals. Support for this working paper series was provided by the authors' home institutions; by GESIS - Leibniz Institute for the Social Sciences; by Schloss Dagstuhl - Leibniz Center for Informatics; and by the DDI Alliance.

Metadata for the Longitudinal Data Life Cycle

THE ROLE AND BENEFIT OF METADATA MANAGEMENT AND REUSE

PROBLEM STATEMENT/DESCRIPTION:

This paper focuses on the unique characteristics of longitudinal studies in which the generation of data and metadata is repeated over time. These types of studies might involve multiple waves, either for a person or a population, or might involve ongoing continuous data collection. Some of the issues that are unique to longitudinal studies follow from the repetitive nature of their data collection. Other issues arise simply due to the extended period over which they are conducted, leaving more opportunity for unanticipated events.

It is important to realize that studies which are not initially intended to be longitudinal may evolve into longitudinal studies. It is therefore best practice for all studies to structure initial metadata to be compatible with this potential repurposing across the data life cycle. Each stage in the workflow may be of particular interest to different groups.

Note: In this document words in italics denote DDI elements, e.g., *StudyUnit*. Also, note that the term “published” when referring to a DDI entity means a DDI instance that has been made available for use outside of the immediate group of its creators. This is denoted by the “isPublished” attribute being set to “true”, and carries with it the requirement that versioning be begun.

APPROACH:

We first listed a number of possible forms longitudinal studies might take:

- Surveillance, data collected through observation over time
- Event-driven data collection
- Panel studies / cohort studies, open cohort studies
- Retrospective studies (probably not “longitudinal”, unless collected at multiple time periods)
- Interventions or trials
- Repeated cross-sections

From this list we chose open cohort studies, one of the more complex designs, as our exemplar, with the thinking that challenges for simpler designs would also be present for the more complex design. We also decided to discuss potential issues in life cycle order as described in Figure 1 below. We wanted to explore what is of particular importance with respect to the temporal aspect of the data. We also drew the distinction

between longitudinal use of the data and longitudinal management of repeated passes through the life cycle stages.

To set the context, the DDI basic life cycle is diagrammed below, with DDI modules connected to the stages of the life cycle for which they are most relevant. For longitudinal studies, other arrows exist in a somewhat different life cycle (see examples in the Repurposing and Redesign section of this paper).

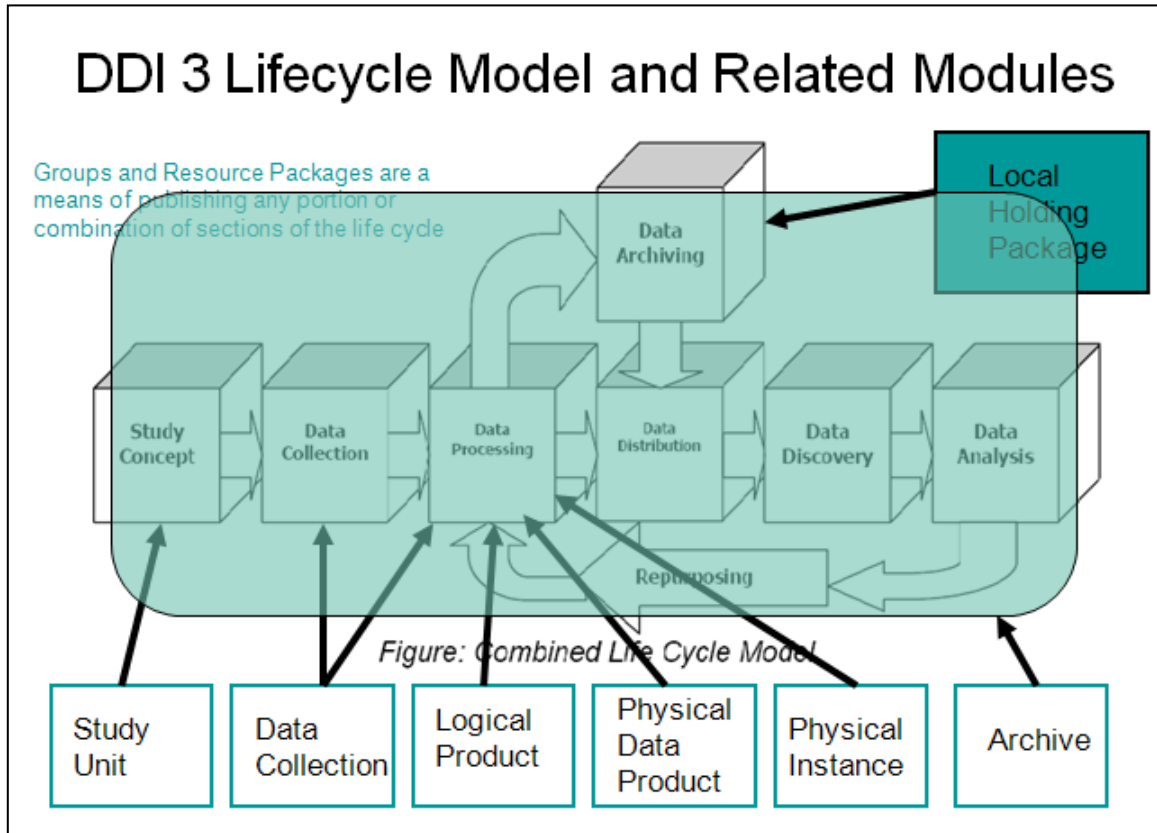


Figure 1: From: *Just Enough DDI 3.ppt*, Arofan Gregory - Dagstuhl 2010 Longitudinal Data Workshop 10422

Study Concept

Universes

Longitudinal studies require careful documentation of a number of aspects throughout their course. The initial study population and study concepts should be described and then described again as they change over time. Sampling procedures, including any number boosting procedures, should be documented thoroughly. The ability to generate an accurate description of the study universe at any given time is essential. The metadata should allow for retrieval of data belonging to any version of a universe or sub-universe from the successive stages of the study.

If the universe expands or changes in any way, the corresponding *Universe* element must have its version updated. Changes in instruments over time may generate changes in universes. An example might be a revision of a skip pattern, causing a question to be answered by a different population. Partitions of the universe may be described hierarchically as *Universe* elements within the parent *Universe* element. See section 3.3 Universe of the DDI 3.1 User Guide for a description of hierarchical universe structures.

A *Comparison* element should be used to describe differences between universe versions. The *VersionRationale* may be used to provide a textual description of the changes. The example below shows documentation of a change in the *Universe* element. Note also that with the change in version of the *Universe*, all of its ancestors (*UniverseScheme*, *ConceptualComponent*, *StudyUnit*, and *DDIInstance*) have a version update. A *VersionRationale* is included for each. Also note the use of a *LifecycleEvent* to document the external event and its related change in the universe.

Example 1 shows a change in *Universe* (❶) documented in DDI 3.1. A *Group* element (❷) contains a *Purpose* and a *Comparison*. The *Comparison* contains a *UniverseMap* pointing to the initial version 1.0.0 (*SourceSchemeReference*) (❸) and the updated *Universe*, version 1.1.0 (*TargetSchemeReference*) (❹). The *Correspondence* element (❺) describes *Commonality* and *Difference* between the versions. A *LifecycleEvent* (❻) describes the external event precipitating the change. Note how the *TargetSchemeReference* (❹) includes a reference to both the *Universe* (❶) and its parent maintainable (*UniverseScheme*).

Example 1 – A Change in Universe

```
<?xml version="1.0" encoding="UTF-8"?>
<ddi:DDIInstance xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xmlns:dc="ddi:dcelements:3_1"
  xmlns:g="ddi:group:3_1"
  xmlns:d="ddi:datacollection:3_1"
  xmlns:c="ddi:conceptualcomponent:3_1"
  xmlns:xhtml="http://www.w3.org/1999/xhtml"
  xmlns:a="ddi:archive:3_1"
  xmlns:m1="ddi:physicaldataproduc_ncube_normal:3_1"
  xmlns:m2="ddi:physicaldataproduc_ncube_tabular:3_1"
  xmlns:ddi="ddi:instance:3_1"
  xmlns:m3="ddi:physicaldataproduc_ncube_inline:3_1"
  xmlns:l="ddi:logicalproduct:3_1"
  xmlns:m4="ddi:physicaldataproduc_proprietary:3_1"
  xmlns:dc2="http://purl.org/dc/elements/1.1/"
  xmlns:cm="ddi:comparative:3_1"
  xmlns:s="ddi:studyunit:3_1"
  xmlns:r="ddi:reusable:3_1"
  xmlns:p="ddi:physicaldataproduc:3_1"
  xmlns:pi="ddi:physicalinstance:3_1"
  xmlns:pr="ddi:ddiprofile:3_1"
  xmlns:ds="ddi:dataset:3_1"
  xsi:schemaLocation="ddi:instance:3_1 instance.xsd"
  id="example_UniverseChange_DDIInstance"
  agency="us.example"
  version="1.1.0"
  versionDate="2010-10-31">
  <r:VersionRationale>Universe updated to version 1.1.0</r:VersionRationale>
  <g:Group id="example_UniverseChange_Comparisons" agency="us.example" version="1.0.0" versionDate="2010-10-31">
    <g:Purpose id="example_UniverseChange_Comparisons_Purpose">
      <r:Content>A group to contain comparisons for changes</r:Content>
    </g:Purpose>
    <cm:Comparison id="example_UniverseChange_TOv1_1_0" agency="us.example" version="1.0.0" versionDate="2010-10-31">
      <cm:UniverseMap id="example_UniverseChange_UMap" version="1.0.0" versionDate="2010-10-31">
        <cm:SourceSchemeReference> (not shown)
          <r:Scheme>
            <r:ID>example_UniverseChange_UniverseScheme</r:ID>
            <r:IdentifyingAgency>us.example</r:IdentifyingAgency>
            <r:Version>1.0.0</r:Version>
          </r:Scheme>
          <r:ID>example_UniverseChange_FrenchWorkers</r:ID>
          <r:IdentifyingAgency>us.example</r:IdentifyingAgency>
          <r:Version>1.0.0</r:Version>
        </cm:SourceSchemeReference>
      </cm:UniverseMap>
    </cm:Comparison>
  </g:Group>
</ddi:DDIInstance>
```

```
</cm:SourceSchemeReference>
```

```
<cm:TargetSchemeReference>
  <r:Scheme>
    <r:ID>example_UniverseChange_UniverseScheme</r:ID>
    <r:IdentifyingAgency>us.example</r:IdentifyingAgency>
    <r:Version>1.1.0</r:Version>
  </r:Scheme>
  <r:ID>example_UniverseChange_FrenchWorkers</r:ID>
  <r:IdentifyingAgency>us.example</r:IdentifyingAgency>
  <r:Version>1.1.0</r:Version>
</cm:TargetSchemeReference>
```

4

```
<cm:Correspondence>
```

```
<cm:Commonality>Both versions are intended to cover people of pre-retirement age in France.</cm:Commonality>
<cm:Difference>In 2010 the French Senate voted to raise the retirement age to 62.</cm:Difference>
```

5

```
</cm:Correspondence>
```

```
</cm:UniverseMap>
</cm:Comparison>
</g:Group>
```

```
<s:StudyUnit id="example_UniverseChange_StudyUnit" agency="us.example" version="1.1.0" versionDate="2010-10-31">
```

```
<r:VersionRationale>Universe updated to version 1.1.0</r:VersionRationale>
```

```
<r:Citation>
```

```
<r:Title>Example Study Unit</r:Title>
```

```
</r:Citation>
```

```
<s:Abstract id="example_UniverseChange_StudyUnit_Abstract">
```

```
<r:Content>An example for:
```

Best Practices: Metadata for the Longitudinal Data Life cycle

The Role and Benefit of Metadata Management and Reuse.

This example demonstrates the documentation of a change in a universe.

```
</r:Content>
```

```
</s:Abstract>
```

```
<r:UniverseReference>
```

```
<r:Scheme>
```

```
<r:ID>example_UniverseChange_UniverseScheme</r:ID>
```

```
<r:IdentifyingAgency>us.example</r:IdentifyingAgency>
```

```
<r:Version>1.1.0</r:Version>
```

```
</r:Scheme>
```

```
<r:ID>example_UniverseChange_FrenchWorkers</r:ID>
```

```
<r:IdentifyingAgency>us.example</r:IdentifyingAgency>
```

```
<r:Version>1.1.0</r:Version>
```

```
</r:UniverseReference>
```

```
<s:Purpose id="example_UniverseChange_Purpose">
```

```
<r:Content>An example of documenting the change in a Universe description</r:Content>
```

```
</s:Purpose>
```

```
<c:ConceptualComponent id="example_UniverseChange_ConceptualComponent" version="1.1.0" versionDate="2010-10-31"
agency="us.example">
```

```
<r:VersionRationale>Universe updated to version 1.1.0</r:VersionRationale>
```

```
<c:UniverseScheme id="example_UniverseChange_UniverseScheme" version="1.1.0" versionDate="2010-10-31"
agency="us.example">
```

```
<r:VersionRationale>Universe updated to version 1.1.0</r:VersionRationale>
```

```
<c:UniverseSchemeName xml:lang="en-US"></c:UniverseSchemeName>
```

```
<c:Universe id="example_UniverseChange_FrenchWorkers" version="1.1.0" versionDate="2010-10-31">
```

```
<r:VersionRationale>In 2010 the French Senate voted to raise the retirement age to 62 from 60.</r:VersionRationale>
```

```
<r:Label xml:lang="en-US">People of Working age in France</r:Label>
```

```
<c:HumanReadable xml:lang="en-US">People of Working age in France. Note that in 2010 the French Senate voted to raise the
retirement age to 62 from 60. The Universe then included more people and the mean age of those in the universe
increased.</c:HumanReadable>
```

```
</c:Universe>
```

```
</c:UniverseScheme>
```

```
</c:ConceptualComponent>
```

```
<a:Archive id="example_UniverseChange_Archive" agency="us.example" version="1.0.0" versionDate="2010-10-31">
```

```
<a:ArchiveSpecific>
```

```
<a:ArchiveOrganizationReference>
```

1

```

<r:Scheme>
  <r:ID>example_UniverseChange_OrgSch</r:ID>
  <r:IdentifyingAgency>us.example</r:IdentifyingAgency>
  <r:Version>1.1.0</r:Version>
</r:Scheme>
<r:ID>example_UniverseChange_Org</r:ID>
<r:IdentifyingAgency>us.example</r:IdentifyingAgency>
<r:Version>1.1.0</r:Version>
</a:ArchiveOrganizationReference>
</a:ArchiveSpecific>
<a:OrganizationScheme id="example_UniverseChange_OrgSch" agency="us.example" version="1.0.0" versionDate="2010-10-31">
  <a:Organization id="example_UniverseChange_Org" version="1.0.0" versionDate="2010-10-31">
    <a:OrganizationName>EXAMPLE</a:OrganizationName>
    <r:Description>An imaginary organization used for examples</r:Description>
  </a:Organization>
</a:OrganizationScheme>
<r:LifecycleInformation>
  <r:LifecycleEvent id="example_UniverseChange_Uch1">
    <r:Date><r:SimpleDate>2010-10-27</r:SimpleDate></r:Date>
    <r:AgencyOrganizationReference> 
    <r:Scheme>
      <r:ID>example_UniverseChange_OrgSch</r:ID>
      <r:IdentifyingAgency>us.example</r:IdentifyingAgency>
      <r:Version>1.1.0</r:Version>
    </r:Scheme>
    <r:ID>example_UniverseChange_Org</r:ID>
    <r:IdentifyingAgency>us.example</r:IdentifyingAgency>
    <r:Version>1.1.0</r:Version>
    </r:AgencyOrganizationReference>
    <r:Description>The French Senate and National Assembly voted to raise the retirement age to 62. This changes the universe of working
age people.</r:Description>
  </r:LifecycleEvent>
</r:LifecycleInformation>
</a:Archive>
</s:StudyUnit>
</ddi:DDIInstance>

```

7

6

Versioning

The extended time frame of longitudinal studies makes changes to metadata likely. These changes necessitate versioning of “published” metadata. The strategy used for documenting versioning should be carefully described at the outset. This will be especially important with DDI 3.2 with its more flexible versioning notation. (Note that the versioning rules and version format have changed with DDI 3.1 and will change again with DDI 3.2.)


In Example 2, a *ResourcePackage* (❶) contains the *Organization* (❷) information to be used by reference. The *Organization* element contains a *Note* (❸) element outlining the versioning structure as recommended by the Best Practices paper on versioning (see DDI Working Paper Series -- Best Practices, No. 8, <http://dx.doi.org/10.3886/DDIBestPractices08>, based on DDI 3.0). That paper also points out the importance of *versionDate* and *VersionRationale* and indicates that versioning events should be documented in *LifecycleEvents* as seen above in Example 1. The *Content* (❹) of the *Note* documents the particular organization’s unique rubric for versioning. Each organization may have its own rules for doing versioning.

Example 2 – Documenting the Versioning Method

```

<?xml version="1.0" encoding="UTF-8"?>
<ddi:DDIInstance xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
[DDI XML namespace definitions suppressed, for details see example 1]
xsi:schemaLocation="ddi:instance:3_1 instance.xsd"
id="AboutOrganization_DDIInstance"
agency="us.example"
version="1.0.0"
versionDate="2010-10-31">

  <g:ResourcePackage id="AboutOrganization" agency="us.example" version="1.0.0" versionDate="2010-10-31"> ❶
    <g:Purpose id="AboutOrganization_Purpose">
      <r:Content>This resource package contains common information about the example organization</r:Content>
    </g:Purpose>

    <a:OrganizationScheme id="AboutOrganization_OrgSch" agency="us.example" version="1.0.0" versionDate="2010-10-31">
      <a:Organization id="AboutOrganization_Org" version="1.0.0" versionDate="2010-10-31"> ❷
        <a:OrganizationName>Example</a:OrganizationName>
        <r:Description>Us.example is an organization name used for documentation</r:Description>
        <r:Note type="Addendum" id="example_UniverseChange_VersioningNote" > ❸
          <r:Relationship>
            <r:RelatedToReference>  ❷
            <r:Scheme>
              <r:ID>AboutOrganization_OrgSch</r:ID>
              <r:IdentifyingAgency>us.example</r:IdentifyingAgency>
              <r:Version>1.0.0</r:Version>
            </r:Scheme>
            <r:ID>AboutOrganization_Org</r:ID>
            <r:IdentifyingAgency>us.example</r:IdentifyingAgency>
            <r:Version>1.0.0</r:Version>
          </r:Relationship>
          <r:Content>The versioning scheme in this DDIInstance is as follows: ❹
            All versions will consist of a string with three numbers separated by periods.
            Major changes produce an increment in the number to the left of the first period.
            Minor but meaningful changes produce an increment in the number between the decimal points.
            Very minor changes, such as correction of typographical errors, produce an increment in the number to the right of the second
            period. Late binding is not used. Elements will be marked as “isPublished” when the DDI is posted to the public site.
            versionDate is updated on unpublished metadata. Our initial version is always 1.0.0.
          </r:Content>
        </r:Note>
      </a:Organization>
    </a:OrganizationScheme>
  </g:ResourcePackage>
</ddi:DDIInstance>

```

Concepts

Concepts may also evolve as the study progresses. Structured concepts may be useful in longitudinal studies. They can be created with a hierarchy of *ConceptGroup* (using references to single concepts) in *ConceptualComponent*. Nesting of *Concepts* within *Concepts* will be available in DDI 3.2. An example from a demographic surveillance site (DSS) study like the INDEPTH Network would be “household at location”, then “social group”, where the concept of social groups refines over time (INDEPTH Network site: <http://www.indepth-ishare.org/>).

Best practice in documenting concepts is to use an existing controlled vocabulary, a thesaurus, or a DDI *ResourcePackage* when they exist. For more see Jääskeläinen et al.

In Example 3, two *Concepts*, “social conservative” (❶) and “economic conservative” (❷), are grouped in a higher level *ConceptGroup* – “conservative” (❸).

Example 3 – Structured Concepts

```
<?xml version="1.0" encoding="UTF-8"?>
<ddi:DDIInstance xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
[DDI XML namespace definitions suppressed, for details see example 1]
xsi:schemaLocation="ddi:instance:3_1 instance.xsd"
id="StructuredConcepts_DDIInstance"
agency="us.example"
version="1.1.0"
versionDate="2010-10-31">
  <g:ResourcePackage id="StructuredConcepts" agency="us.example" version="1.0.0" versionDate="2010-10-31">
    <g:Purpose id="StructuredConcepts_Purpose">
      <r:Content>This Resource Package contains example structured concepts</r:Content>
    </g:Purpose>
    <c:ConceptScheme id="StructuredConcepts_ConceptScheme" agency="us.example" version="1.0.0" versionDate="2010-10-31">
      <c:Concept id="StructuredConcepts_CG_conservativeSocial" version="1.0.0" versionDate="2010-10-31">❶
        <c:ConceptName>SocialConservative</c:ConceptName>
        <r:Label>Self identified as socially conservative.</r:Label>
        <r:Description>Persons identifying themselves as socially conservative</r:Description>
        <c:SimilarConcept>
          <c:SimilarConceptReference>❸❷
            <r:Scheme>
              <r:ID>StructuredConcepts_ConceptScheme</r:ID>
              <r:IdentifyingAgency>us.example</r:IdentifyingAgency>
              <r:Version>1.0.0</r:Version>
            </r:Scheme>
            <r:ID>StructuredConcepts_CG_conservativeEconomic</r:ID>
            <r:IdentifyingAgency>us.example</r:IdentifyingAgency>
            <r:Version>1.0.0</r:Version>
          </c:SimilarConceptReference>
        </c:SimilarConcept>
      </c:Concept>
      <c:Concept id="StructuredConcepts_CG_conservativeEconomic" version="1.0.0" versionDate="2010-10-31">❷
        <c:ConceptName>EconomicConservative</c:ConceptName>
        <r:Label>Self identified as economically conservative</r:Label>
        <r:Description>Persons identifying themselves as economically conservative</r:Description>
        <c:SimilarConcept>
          <c:SimilarConceptReference>❸❶
            <r:Scheme>
              <r:ID>StructuredConcepts_ConceptScheme</r:ID>
              <r:IdentifyingAgency>us.example</r:IdentifyingAgency>
              <r:Version>1.0.0</r:Version>
            </r:Scheme>
          </c:SimilarConceptReference>
        </c:SimilarConcept>
      </c:Concept>
    </c:ConceptScheme>
  </g:ResourcePackage>
</ddi:DDIInstance>
```



```

</r:Scheme>
<r:ID>StructuredConcepts_CG_conservativeSocial</r:ID>
<r:IdentifyingAgency>us.example</r:IdentifyingAgency>
<r:Version>1.0.0</r:Version>
</c:SimilarConceptReference>
</c:SimilarConcept>
</c:Concept>
<c:ConceptGroup purpose="Conceptual" id="StructuredConcepts_CG_conservative" version="1.0.0" versionDate="2010-10-31" > ③
<c:ConceptGroupName>Conservative</c:ConceptGroupName>
<r:Label>Self identified as conservative</r:Label>
<r:Description>Persons identifying themselves as conservative</r:Description>
<c:ConceptReference> ↗ ①
<r:Scheme>
<r:ID>StructuredConcepts_ConceptScheme</r:ID>
<r:IdentifyingAgency>us.example</r:IdentifyingAgency>
<r:Version>1.0.0</r:Version>
</r:Scheme>
<r:ID>StructuredConcepts_CG_conservativeSocial</r:ID>
<r:IdentifyingAgency>us.example</r:IdentifyingAgency>
<r:Version>1.0.0</r:Version>
</c:ConceptReference>
<c:ConceptReference> ↗ ②
<r:Scheme>
<r:ID>StructuredConcepts_ConceptScheme</r:ID>
<r:IdentifyingAgency>us.example</r:IdentifyingAgency>
<r:Version>1.0.0</r:Version>
</r:Scheme>
<r:ID>StructuredConcepts_CG_conservativeEconomic</r:ID>
<r:IdentifyingAgency>us.example</r:IdentifyingAgency>
<r:Version>1.0.0</r:Version>
</c:ConceptReference>
</c:ConceptGroup>
</c:ConceptScheme>
</g:ResourcePackage>
</ddi:DDIInstance>

```

Study Unit

A *Group* structure can enable representing metadata common to waves -- for an example using DDI 3.0, see Goebel and Wackerow 2007. Metadata describing the relationship of *StudyUnits* may be placed in a parent *Group*. A *SeriesStatement* may also be used to document relationships among waves. Metadata to be shared among *StudyUnits* are better represented in a *ResourcePackage*. The consensus was that a *ResourcePackage* is the more machine-actionable structure.

Example 4 shows a *SeriesStatement* (❶) pointing to the series to which its *StudyUnit* (❷) belongs. Any additional *StudyUnit* in the series would contain a similar *SeriesStatement*.

Example 4 – Using a Series Statement

```
<?xml version="1.0" encoding="UTF-8"?>
<ddi:DDIInstance xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
[DDI XML namespace definitions suppressed, for details see example 1]
xsi:schemaLocation="ddi:instance:3_1 instance.xsd"
id="example_SeriesStatement_DDIInstance"
agency="us.example"
version="1.1.0"
versionDate="2010-10-31">

  <s:StudyUnit id="example_SeriesStatement_StudyUnit1" agency="us.example" version="1.1.0" versionDate="2010-10-31"> ❷
    <r:Citation>
      <r:Title>Example Study Unit</r:Title>
    </r:Citation>
    <s:Abstract id="example_SeriesStatement_StudyUnit_Abstract">
      <r:Content>This example demonstrates the use of a SeriesStatement.
    </r:Content>
    </s:Abstract>
    <r:UniverseReference>
      <r:Scheme>
        <r:ID>example_SeriesStatement_UniverseScheme</r:ID>
        <r:IdentifyingAgency>us.example</r:IdentifyingAgency>
        <r:Version>1.1.0</r:Version>
      </r:Scheme>
      <r:ID>example_SeriesStatement_FrenchWorkers</r:ID>
      <r:IdentifyingAgency>us.example</r:IdentifyingAgency>
      <r:Version>1.1.0</r:Version>
    </r:UniverseReference>

    <r:SeriesStatement> ❶
      <r:SeriesRepositoryLocation>us.example/LongitudinalLifecycle</r:SeriesRepositoryLocation> (external)
      <r:SeriesName>Longitudinal Life Cycle Example Study Units</r:SeriesName>
      <r:Abbreviation>LongLifeStudies</r:Abbreviation>
      <r:SeriesDescription>This is a hypothetical series of study units for the Longitudinal Life Cycle Metadata
paper</r:SeriesDescription>
    </r:SeriesStatement>
    <s:Purpose id="example_SeriesStatement_Purpose">
      <r:Content>An example of documenting the use of a SeriesStatement </r:Content>
    </s:Purpose>
    <c:ConceptualComponent id="example_SeriesStatement_ConceptualComponent" version="1.1.0" versionDate="2010-10-31"
agency="us.example">
      <c:UniverseScheme id="example_SeriesStatement_UniverseScheme" version="1.1.0" versionDate="2010-10-31" agency="us.example">
        <r:VersionRationale>Universe updated to version 1.1.0</r:VersionRationale>
        <c:UniverseSchemeName xml:lang="en-US"></c:UniverseSchemeName>
        <c:Universe id="example_SeriesStatement_FrenchWorkers" version="1.1.0" versionDate="2010-10-31">
          <r:VersionRationale>In 2010 the French Senate voted to raise the retirement age to 62 from 60</r:VersionRationale>
          <r:Label xml:lang="en-US">People of Working age in France</r:Label>
          <c:HumanReadable xml:lang="en-US">People of Working age in France. Note that in 2010 the French Senate voted to raise the
retirement age to 62 from 60. The Universe then included more people and the mean age of those in the universe
increased.</c:HumanReadable>
        </c:Universe>
      </c:UniverseScheme>
    </c:ConceptualComponent>
  </s:StudyUnit>
</ddi:DDIInstance>
```

Data Collection

A longitudinal study is particularly well suited to metadata sharing practices. These include management and reuse of *Instruments*, *QuestionSchemes*, and *CollectionEvents* – including *ModeOfCollection*. An overall explicitly described versioning practice will become important here as well as the careful use of *Comparison*.

The following example shows the documentation of a change in a question from *QuestionItem* version 1.0.0 (❶) to *QuestionItem* version 2.0.0 (❷) and their associated categories, *CategoryScheme* version 1.0.0 (❸) and *CategoryScheme* version 2.0.0 (❹). A picture is also added to the revised version of the question. A *QuestionMap* (❺), *CategoryMap* (❻) and *ItemMap* (❼) document the changes, both commonalities and differences.

Note that if there were associated codes, *GenerationInstructions* could be used to describe the relationship of values from the revised version to the original values in a machine-actionable way. No such facility seems to exist for categories. The proper value for a *CommonalityWeight* for changes in the “Other” (❸) & (❹) category in this case is unclear.

Two *ControlConstructSchemes* (❿) are also included to show how a sequence of questions is instantiated. Note that the version numbers within the *ControlConstructSchemes* are not required to match the version numbers of the questions they reference.

Example 5– When Questions Change

```
<?xml version="1.0" encoding="UTF-8"?>
<ddi:DDIInstance xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
[DDI XML namespace definitions suppressed, for details see example 1]
xsi:schemaLocation="ddi:instance:3_1 instance.xsd"
id="QuestionChange_DDIInstance"
agency="us.example"
version="2.0.0"
versionDate="2010-11-04">
<g:ResourcePackage id="QuestionChange" agency="us.example"
version="2.0.0" versionDate="2010-11-04">
<g:Purpose id="QuestionChange_Purpose">
<r:Content>This Resource package contains an example of a change in a question</r:Content>
</g:Purpose>

<cm:Comparison id="QuestionChange_Q_Comparison" agency="us.example" version="1.0.0" versionDate="2010-11-04">
<r:Label>Compares v1.0.0 and v2.0.0 of QuestionChange_CatSch</r:Label>
<r:Description>Compares v1.0.0 and v2.0.0 of QuestionChange_QScheme and QuestionChange_CatSch. Version 2.0.0 added the category
"Brush"</r:Description>

<cm:QuestionMap id="QuestionChange_Q_QMap" version="1.0.0">
<cm:SourceSchemeReference>
<r:ID>QuestionChange_QScheme</r:ID>
<r:IdentifyingAgency>us.example</r:IdentifyingAgency>
<r:Version>1.0.0</r:Version>
</cm:SourceSchemeReference>

<cm:TargetSchemeReference>
```

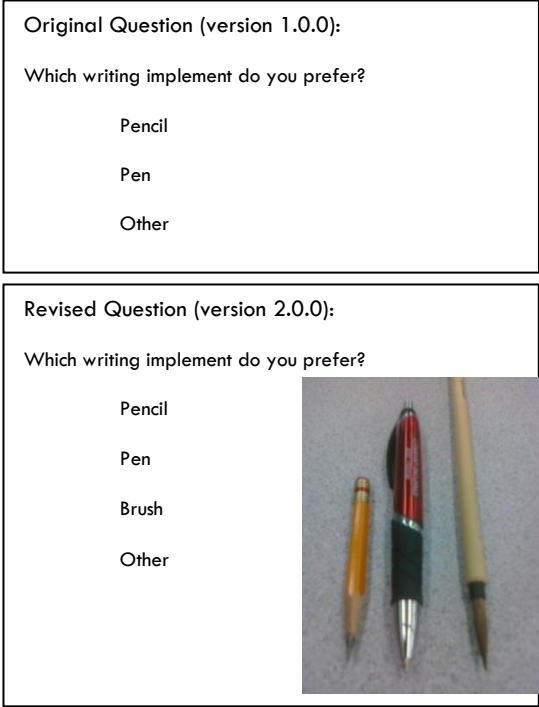




Figure 2: Example Questions

```
<r:ID>QuestionChange_QScheme</r:ID>
<r:IdentifyingAgency>us.example</r:IdentifyingAgency>
<r:Version>2.0.0</r:Version>
</cm:TargetSchemeReference>
```

```
<cm:Correspondence>
  <cm:Commonality></cm:Commonality>
  <cm:Difference>QuestionChange_Q changed</cm:Difference>
</cm:Correspondence>
```

```
<cm:ItemMap>
```

```
<cm:SourceItem>QuestionChange_Q</cm:SourceItem> 
<cm:TargetItem>QuestionChange_Q</cm:TargetItem> 
```

```
<cm:Correspondence>
```

```
  <cm:Commonality>Categories from version 2.0.0 can be aggregated to match version 1.0.0</cm:Commonality>
```

```
  <cm:Difference>Version 2 refers to a revised version of the category scheme QuestionChange_CatSch, which adds
the category Brush. Version 2.0.0 also included an external image.</cm:Difference>
```

```
  <cm:CommonalityWeight>0.9</cm:CommonalityWeight>
```

```
</cm:Correspondence>
```

```
</cm:ItemMap>
```

```
</cm:QuestionMap>
```

```
<cm:CategoryMap id="QuestionChange_CatSch_CatMap" version="1.0.0" versionDate="2010-11-04">
```

6

```
<cm:SourceSchemeReference> 
```

```
<r:ID>QuestionChange_CatSch</r:ID>
```

```
<r:IdentifyingAgency>us.example</r:IdentifyingAgency>
```

```
<r:Version>1.0.0</r:Version>
```

```
</cm:SourceSchemeReference>
```

```
<cm:TargetSchemeReference> 
```

```
<r:ID>QuestionChange_CatSch</r:ID>
```

```
<r:IdentifyingAgency>us.example</r:IdentifyingAgency>
```

```
<r:Version>2.0.0</r:Version>
```

```
</cm:TargetSchemeReference>
```

```
<cm:Correspondence>
```

```
  <cm:Commonality>Both contain the categories Pencil, Pen, and Other
```

```
</cm:Commonality>
```

```
  <cm:Difference>Version 2.0.0 adds the category Brush</cm:Difference>
```

```
</cm:Correspondence>
```

```
<cm:ItemMap>
```

```
<cm:SourceItem>QuestionChange_Pencil</cm:SourceItem>
```

```
<cm:TargetItem>QuestionChange_Pencil</cm:TargetItem>
```

```
<cm:Correspondence>
```

```
  <cm:Commonality>Same</cm:Commonality>
```

```
  <cm:Difference></cm:Difference>
```

```
  <cm:CommonalityWeight>1</cm:CommonalityWeight>
```

```
</cm:Correspondence>
```

```
</cm:ItemMap>
```

```
<cm:ItemMap>
```

```
<cm:SourceItem>QuestionChange_Pen</cm:SourceItem>
```

```
<cm:TargetItem>QuestionChange_Pen</cm:TargetItem>
```

```
<cm:Correspondence>
```

```
  <cm:Commonality>Same</cm:Commonality>
```

```
  <cm:Difference></cm:Difference>
```


```
  <cm:CommonalityWeight>1</cm:CommonalityWeight>
```

```
</cm:Correspondence>
```

```
</cm:ItemMap>
```

```
<cm:ItemMap>
```

```
<cm:SourceItem>QuestionChange_Other</cm:SourceItem> 
```

```
<cm:TargetItem>QuestionChange_Other</cm:TargetItem> 
```

```
<cm:Correspondence>
```

7

```

    <cm:Commonality></cm:Commonality>
    <cm:Difference>Version 1.0.0 category of Other was divided into Other and Brush in version 2.0.0</cm:Difference>
    <cm:CommonalityWeight>0</cm:CommonalityWeight>
  </cm:Correspondence>
</cm:ItemMap>
</cm:CategoryMap>
</cm:Comparison>

```

```

<c:ConceptScheme id="QuestionChange_ConceptScheme" agency="us.example" version="1.0.0" versionDate="2010-10-31">
  <c:Concept id="QuestionChange_PreferredImplement" version="1.0.0" versionDate="2010-10-31">
    <c:ConceptName>PreferredWritingImplement</c:ConceptName>
    <r:Label>Self identified preferred writing implement</r:Label>
    <r:Description>The writing implement preferred</r:Description>
  </c:Concept>
</c:ConceptScheme>

```

10

```

<d:ControlConstructScheme id="QuestionChange_ControlConstructScheme" agency="us.example" version="1.0.0" versionDate="2010-10-31">
  <d:QuestionConstruct id="QuestionConstruct_Q" version="1.0.0">
    <d:QuestionReference>
      <r:URN>URN:DDI: us.example:QuestionScheme.QuestionChange_QScheme.1.0.0:QuestionItem.QuestionChange_Q.1.0.0</r:URN>
    </d:QuestionReference>
  </d:QuestionConstruct>
  <d:QuestionConstruct id="QuestionConstruct_Age" version="1.0.0">
    <d:QuestionReference>
      <r:URN>URN:DDI: us.example:QuestionScheme.QuestionChange_QScheme.1.0.0:QuestionItem.Age.1.0.0</r:URN>
    </d:QuestionReference>
  </d:QuestionConstruct>
</d:ControlConstructScheme>

```

```

<d:ControlConstructScheme id="QuestionChange_ControlConstructScheme" agency="us.example" version="2.0.0" versionDate="2010-11-04">
  <d:QuestionConstruct id="QuestionConstruct_Q" version="2.0.0">
    <d:QuestionReference>
      <r:URN>URN:DDI: us.example:QuestionScheme.QuestionChange_QScheme.2.0.0:QuestionItem.QuestionChange_Q.2.0.0</r:URN>
    </d:QuestionReference>
  </d:QuestionConstruct>
  <d:QuestionConstruct id="QuestionConstruct_Age" version="1.0.0">
    <d:QuestionReference>
      <r:URN>URN:DDI: us.example:QuestionScheme.QuestionChange_QScheme.1.0.0:QuestionItem.Age.1.0.0</r:URN>
    </d:QuestionReference>
  </d:QuestionConstruct>
</d:ControlConstructScheme>

```

```

<d:QuestionScheme id="QuestionChange_QScheme" agency="us.example" version="1.0.0" versionDate="2010-11-04">
  <d:QuestionSchemeName>Example Question scheme</d:QuestionSchemeName>
  <r:Label>QuestionScheme for an example</r:Label>
  <r:Description>A QuestionScheme for an example of a change in a question</r:Description>

```

```

<d:QuestionItem id="QuestionChange_Q" version="1.0.0" versionDate="2010-10-31">
  <d:QuestionText>
    <d:LiteralText>
      <d:Text>Which writing implement do you prefer?</d:Text>
    </d:LiteralText>
  </d:QuestionText>
  <d:CategoryDomain>

```

1

```

    <r:CategorySchemeReference>
      <r:ID>QuestionChange_CatSch</r:ID>
      <r:IdentifyingAgency>us.example</r:IdentifyingAgency>
      <r:Version>1.0.0</r:Version>
    </r:CategorySchemeReference>
  </d:CategoryDomain>
  <d:ConceptReference>
    <r:Scheme>
      <r:ID>QuestionChange_ConceptScheme</r:ID>

```

```

    <r:IdentifyingAgency>us.example</r:IdentifyingAgency>
    <r:Version>1.0.0</r:Version>
  </r:Scheme>
  <r:ID>QuestionChange_PreferredImplement</r:ID>
  <r:IdentifyingAgency>us.example</r:IdentifyingAgency>
  <r:Version>1.0.0</r:Version>
</d:ConceptReference>
</d:QuestionItem>

```

```

<d:QuestionItem id="Age" version="1.0.0">
  <d:QuestionText>
    <d:LiteralText>
      <d:Text>How Old are You?</d:Text>
    </d:LiteralText>
  </d:QuestionText>
  <d:NumericDomain type="Integer"></d:NumericDomain>
</d:QuestionItem>
</d:QuestionScheme>

```

```

<d:QuestionScheme id="QuestionChange_QScheme" agency="us.example" version="2.0.0" versionDate="2010-11-04">
  <d:QuestionSchemeName>Example Question scheme</d:QuestionSchemeName>
  <r:Label>QuestionScheme for an example</r:Label>
  <r:Description>A QuestionScheme for an example of a change in a question</r:Description>

```

```

<d:QuestionItem id="QuestionChange_Q" version="2.0.0" versionDate="2010-11-04">
  <r:VersionRationale>Changed reference to CategoryScheme adding "Brush", added a picture</r:VersionRationale>
  <d:QuestionText>
    <d:LiteralText>
      <d:Text>Which writing implement do you prefer?</d:Text>
    </d:LiteralText>
  </d:QuestionText>
  <d:CategoryDomain>
    <r:CategorySchemeReference>
      <r:ID>QuestionChange_CatSch</r:ID>
      <r:IdentifyingAgency>us.example</r:IdentifyingAgency>
      <r:Version>2.0.0</r:Version>
    </r:CategorySchemeReference>
  </d:CategoryDomain>
  <d:ConceptReference>
    <r:Scheme>
      <r:ID>QuestionChange_ConceptScheme</r:ID>
      <r:IdentifyingAgency>us.example</r:IdentifyingAgency>
      <r:Version>1.0.0</r:Version>
    </r:Scheme>
    <r:ID>QuestionChange_PreferredImplement</r:ID>
    <r:IdentifyingAgency>us.example</r:IdentifyingAgency>
    <r:Version>1.0.0</r:Version>
  </d:ConceptReference>
  <d:ExternalAid type="Photo" id="QuestionChange_Q_Photo">
    <r:Citation>
      <r:Title>Pencil, Pen, and Brush</r:Title>
    </r:Citation>
    <r:ExternalURLReference>http://us.example/PencilPenBrush.jpg</r:ExternalURLReference>
    <r:MIMEType>image/jpeg</r:MIMEType>
  </d:ExternalAid>
</d:QuestionItem>
</d:QuestionScheme>

```

```

<l:CategoryScheme id="QuestionChange_CatSch" agency="us.example" version="1.0.0" versionDate="2010-10-31">
  <r:Label>Writing Implements</r:Label>
  <r:Description>A list of writing implements</r:Description>
  <l:Category id="QuestionChange_Pencil" version="1.0.0">

```

2

3

```

<l:CategoryName>PencilPreferred</l:CategoryName>
<r:Label>Pencil</r:Label>
<r:Description>Pencil</r:Description>
</l:Category>

```

```

<l:Category id="QuestionChange_Pen" version="1.0.0">
  <l:CategoryName>PenPreferred</l:CategoryName>
  <r:Label>Pen</r:Label>
  <r:Description>Pen</r:Description>
</l:Category>

```

```

<l:Category id="QuestionChange_Other" version="1.0.0">
  <l:CategoryName>OtherPreferred</l:CategoryName>
  <r:Label>Other</r:Label>
  <r:Description>Other</r:Description>
</l:Category>
</l:CategoryScheme>

```

8

```

<l:CategoryScheme id="QuestionChange_CatSch" agency="us.example" version="2.0.0" versionDate="2010-11-
04">

```

4

```

  <r:VersionRationale>Added the category of "Brush"</r:VersionRationale>
  <r:Label>Writing Implements</r:Label>
  <r:Description>A list of writing implements</r:Description>
  <l:Category id="QuestionChange_Pencil" version="2.0.0">
    <l:CategoryName>PencilPreferred</l:CategoryName>
    <r:Label>Pencil</r:Label>
    <r:Description>Pencil</r:Description>
  </l:Category>

```

```

  <l:Category id="QuestionChange_Pen" version="2.0.0">
    <l:CategoryName>PenPreferred</l:CategoryName>
    <r:Label>Pen</r:Label>
    <r:Description>Pen</r:Description>
  </l:Category>

```

```

  <l:Category id="QuestionChange_Brush" version="2.0.0">
    <l:CategoryName>BrushPreferred</l:CategoryName>
    <r:Label>Brush</r:Label>
    <r:Description>Brush</r:Description>
  </l:Category>

```

9

```

  <l:Category id="QuestionChange_Other" version="2.0.0">
    <l:CategoryName>OtherPreferred</l:CategoryName>
    <r:Label>Other</r:Label>
    <r:Description>Other</r:Description>
  </l:Category>
</l:CategoryScheme>
</g:ResourcePackage>
</ddi:DDIInstance>

```

Data Processing

Compatibility of *CategorySchemes* and *CodeSchemes* with relevant external data sources should be planned at the outset. This may include both public data sources and specific related target studies. Care should be taken to identify representations compatible with community standards, e.g., standard demographic variables. *Comparison* can be used to document differences from the community standards. *Comparison* should also be used to document version changes across waves or stages of the project. Common *CategorySchemes* and *CodeSchemes* should be used for measuring the same constructs.

Use controlled vocabularies where available. The advantages of using controlled vocabularies in enhancing comparability across studies (especially in a machine-actionable sense) also hold for longitudinal studies where, for example, there are changes in staff across time. The DDI Controlled Vocabularies Working Group is developing a set of controlled vocabularies (see Jääskeläinen et al.).

Document the computational methods and tools for derived variables such as scored scales. Also document the metrics used for quality assurance during the project as well as changes driven by these techniques.

Data Distribution

During the course of a longitudinal study there will be occasions that call for the preservation of the state of the project, or some subset of the project. An example might be the need to preserve or publish the exact data used for a particular published analysis. The metadata associated with those data must also be extracted. A preserved *PhysicalDataProduct* instance will contain references to the appropriate version of referenced metadata for a particular *PhysicalInstance*.

Data Archiving

Various events during the course of the project should be documented with *LifecycleEvents*. One such event would be the generation of a snapshot of the data. A record of distribution of extracts or snapshots could be kept. Other events would include changes in people or organizations associated with the project, funding source changes, change to software used, other infrastructure or technology changes, external events (e.g., health clinic opens, electricity or water become available, geopolitical boundaries change, freedom of information requests are made, national disasters occur), contractor changes, legislation changes, or changes in study focus.

Note that *LifecycleEvents* may be used with a controlled vocabulary. There is currently one being developed for administrative events along the data life cycle. A first set of controlled vocabularies will be published by the DDI Alliance in the first quarter of 2011.

The role individuals play in the project may be documented in *Role* elements within *OrganizationScheme* elements. Changes in consent should also be documented as they occur.

Data Discovery

Using the metadata, users should be able to find linkages across time and or waves. Changes to consent may affect the ability to expose data to users. Current applicable rights must be determinable at any time during the project.

Data Analysis

Metadata should be made available to analysts in a transparent manner. Versioning and Comparisons generated during earlier stages will be important to the analyst.

Repurposing and Redesign

The life cycle for longitudinal studies is somewhat different than that for other studies. Events during the course of the project may provoke later changes at the study concept or data collection stages of the life cycle model. The notion of repurposing becomes part of the project itself as a result of unexpected events, results, and of quality assurance practices. Project data, or derived subsets, may be archived, tied to certain events, such as an analysis for a publication. Best practice is to use ongoing results to improve the study process, while documenting any changes and methods for generating comparable data. Instances of changes making data not comparable must also be carefully documented. Use of the data must take all of this metadata into account.

Figure 3a shows an example life cycle, where the study concept is modified as the study proceeds. Time is shown progressing upward and the stages of the life cycle are depicted as spiraling around an archive. While this diagram shows two discrete waves of data collection, with two collection efforts in each wave, other designs may be more complex, with multiple ongoing, overlapping data collection streams. A really complex study could potentially have simultaneous activity in each of the life cycle stages, redesign of future waves, ongoing data collection, processing, distribution, discovery, analysis, and archiving. Any one of the stages could prompt revisions in other stages.

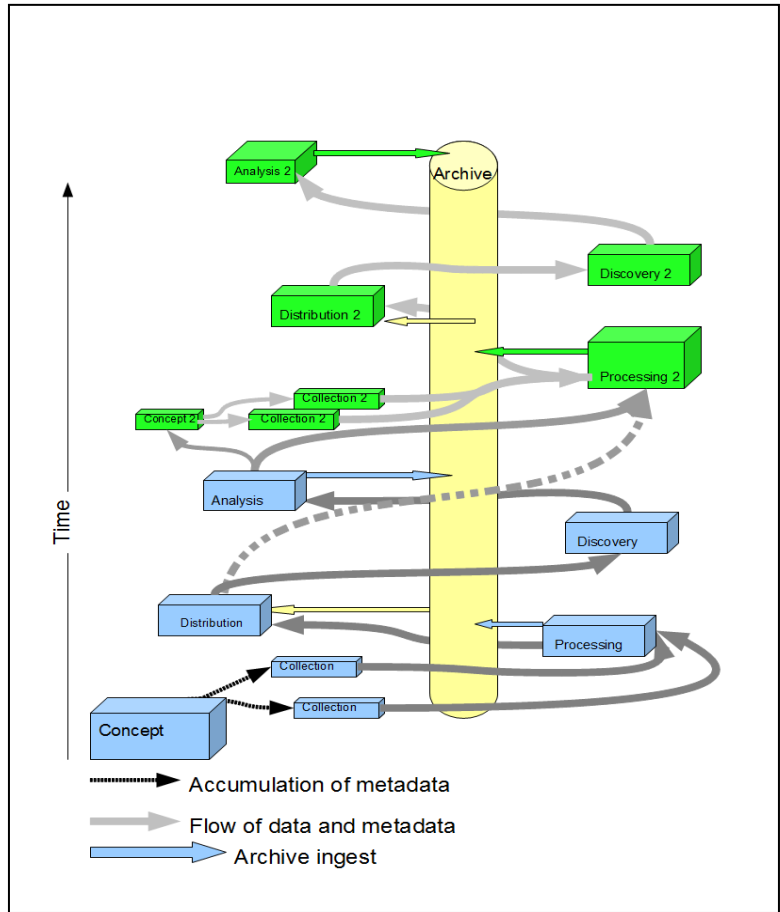


Figure 3a: Example of the longitudinal life cycle – side view

Metadata begin flowing at the initial concept stage, and then a combination of metadata and data flow from the first data collection onward. In this diagram data and metadata flow from the initial analysis to the second processing stage as well as to a second concept stage. In a longitudinal study findings in an earlier stage may result in a reconceptualization of the study. There can be many other connections. Data from an earlier stage may be used in later collection stages. They may be presented to subjects as stimuli, or may affect which items are collected – e.g., a subset of subjects may be asked certain questions based on their answer to questions in an earlier stage.

Figure 3b shows the same diagram from the top, with time spiraling outward. Nothing of the complex processes that goes on inside the archive is shown here. The larger green and blue arrows represent ingest packages from the perspective of the archive. For a life cycle model from the archive perspective see the DCC Curation Life Cycle Model referenced below.

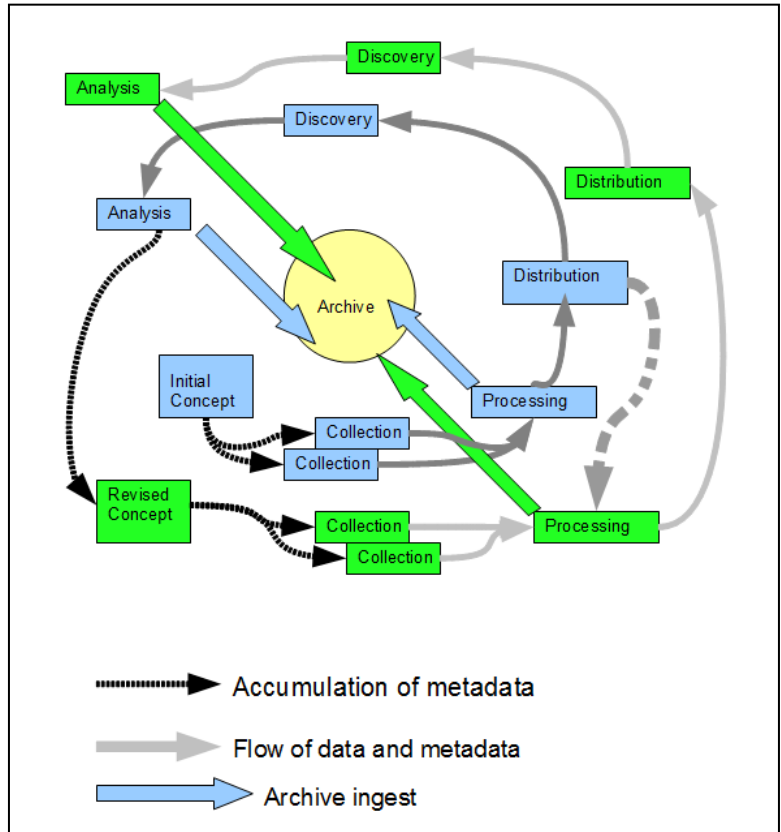


Figure 3b: Example of the longitudinal life cycle – top view

Again, this is a simplified model. It does not show all of the potential linkages where information or events in earlier stages may affect specific later stages. Collection stages may affect later concept stages. Processing stages may reveal the need to change later collection stages. This model also does not directly show the impact of external events on the life cycle. Social, political, and economic events might require changes to the design or concept of the study. Repurposing of the data and metadata by external parties might also induce changes.

RECOMMENDATIONS:

Plan for Change

The extended length of a longitudinal study increases the risk of an unanticipated event necessitating some change in the project.

- Make use of the modularity of DDI.
- Plan for migration across evolving infrastructure. Software and hardware may well change across the course of the study. Document these changes. *LifecycleEvent* in *Archive* is the appropriate element for this documentation.
- Plan for change in technology. A good example of unanticipated change is the relatively recent emergence of biomarkers as important data in longitudinal studies. At the beginning of some of the longer studies, these data were not very practical, both economically and procedurally (requiring blood draws). As these barriers have disappeared, these measures have been increasingly included in ongoing studies. Mid-study additions such as these will require thorough documentation.
- Identifiers should not contain compound information. An example where a compound identification scheme may go awry would be where an identifier consists of village, and person within village. When the person moves to another village the original identifier no longer matches the scheme. A better scheme in this example would be to use separate, unique identifiers for person and village.
- Embedding metadata in variable names can lead to problems. An example would be using a character in the variable name to indicate the wave (e.g., HeightC being height in the third wave). What happens when you run out of unique characters? This is more likely when data are stored in “wide” format. Better practice might be to store data in “narrow” format. In the preceding example there would one variable for height and one for wave.
- Introducing new languages and cultures brings up issues of comparability – document with *Comparison*.
- Plan for how you will deal with changes in the most recent version of DDI itself.

Time

- With any data having a temporal aspect, consider potential future use as part of a longitudinal study. The form of temporal measurement should receive due consideration – the calendar used should be compatible with other anticipated data sources.
- Be clear about the semantics of describing time.

Record Linkage

- Relationships with third-party providers involving record linkage may change over time. Consideration should be made as to how identifiers are managed in these situations.
- Methods for linkage to external sources should be documented. *LogicalRecord* is the appropriate element for this documentation.

Metadata Sharing

- In a complex longitudinal design there may be multiple alternatives for structuring the metadata into *StudyUnit* elements. This choice should be informed by the anticipation of which metadata elements will always be used together. Metadata in separate *StudyUnits* can be shared with *ResourcePackages* or *Groups*.
- Pushing metadata too high in a *Group* hierarchy may cause future effort by requiring add/delete at lower levels. Suppose, for example, two subgroups initially use a *CategoryScheme* by inheritance from a parent group and then one of the subgroups requires the addition of a *Category*. This will require an action attribute to allow that subgroup to pass on the revised metadata to its descendants.
- One goal in designing the sharing of metadata is efficiency - not repeating documentation, and maximizing reuse. Proper sharing design will produce a benefit from previous stages of the life cycle (see Iverson, 2009).
- Another goal should be to improve data and metadata quality over time.
- Use standard controlled vocabularies when possible.

Resource Package or Group?

- Best practice for sharing metadata is to use *ResourcePackage* and use the metadata by reference. This is a more machine-actionable approach and is more flexible than using *Group*.
- Where *Group* is used for sharing the grouping hierarchy should be unlikely to change. A better use of *Group* is for indicating similarity.

Versioning

Managing versioning is an important part of managing a longitudinal study.

- Top-level metadata should contain a detailed description of the process for managing versioning, including:
 - What triggers a new version
 - How you define identity – when identity is “different,” then a version changes.
 - When, if ever, late binding is used.
 - When metadata will be marked in DDI as “published”, which then requires versioning. In general this will be when the metadata become available for use or reference outside of the immediate project team.

Some other points about versioning:

- Once the DDI “isPublished” attribute is set to true, if any property of an element changes, its version changes, and the version of all of its ancestors in the XML hierarchy changes. Note that in DDI 3.1 this leads to a problem with complete identifiers of versionables. In general this issue is that when a versionable changes, the version of its parent changes which, in turn, changes the identifier of all of its unchanged siblings.¹ Example 5 above shows the recommended approach to changes in questions in DDI3.1 - not including the unchanging questions in the revised scheme, and referencing the questions in a *ControlCounstructScheme*.
- In general, late binding is a bad idea, producing the risk that referenced metadata will no longer be appropriate. Metadata for snapshots of a project in particular should not use late binding. Late binding can only make sense when strict rules are established among partners regarding what triggers a new version on which level, i.e. changes caused by correcting spelling errors.
- Two versions of the same thing should not exist in the same wave; nothing should be referring to both at the same time. An exception to this might be where two versions of a question both appear in the same wave.
- Use the *VersionResponsibility* and *VersionRationale* elements along with the version and versionDate attributes when versions change.
- Could the change possibly affect an analysis? If so, document that possible effect, maybe in a *LifecycleEvent*. When a variable changes, use *Comparison* between the versions.
- Consider Change Management best practices when looking at versioning (see, for example, http://en.wikipedia.org/wiki/Change_control)

A Note about Comparison:

Changes which might affect future use should be documented in a *Comparison* element where possible. This is the most machine-actionable practice. As of DDI 3.1 not every element for which a *Comparison* might be generated, for example between *StudyUnit* elements, can be included in a *Comparison*. In those cases a *Note* would be better than nothing.

¹ In mid December 2010, an issue was raised on the ddi-users listserv regarding situations in which a versionable element may end up with multiple distinct identifiers in DDI 3.1. A scenario generating this situation would be that a sibling element in the same scheme is modified. This in turn would require the reversioning of the parent maintainable. The unchanged versionable would then have two valid URNs, one referencing the original (parent) maintainable and the other the new maintainable.

This example, based on a real use case, involved a *QuestionScheme* *qsS*, version 1.0.0 with three *QuestionItems* *qA*, *qB*, and *qC*, all version 1.0.0. If *qC* is modified and becomes version 2.0.0 then *QuestionScheme* *qsS* must have a new version number, say version 2.0.0. *Question* *qA* now can be identified as (*qsS* 1.0.0: *qA* 1.0.0) and (*qsS* 2.0.0: *qA* 1.0.0), with no machine-actionable method to determine that both references refer to the same, unchanged, question.

This issue will be resolved in DDI 3.2. In the interim, with DDI 3.1, best practice is to define the unchanged elements in a revised scheme by reference to their original definitions. This documents their equivalence.

General Practices

- Metadata can be used in the management of the study as well as for documenting it for posterity. Some uses include:
 - Metadata-driven survey design (see Iverson 2009)
 - Integration of existing systems
 - Avoiding writing unnecessary code
 - Metadata-driven external applications
- DDI for a longitudinal study can be substantially more complex than for a simple one-time study. Example 5, a simple change in a question, gives some sense of this. Tools for managing and generating DDI will be important. One example is maintaining unique IDs at the agency level. Some mechanism for ensuring that uniqueness is necessary, either by tracking IDs used or by adopting tools to generate unique IDs through a standard like ISO/IEC 9834-8.

Additions to DDI

- DDI may need a facility for recording the quality of linkage among datasets, and a description of what to do under different quality levels. There are a number of factors that can affect the ability to link external files. Data may be missing in the fields (keys) needed to match records or the keys may be coded differently. Reconciling differences may be a simple matter of a one to one replacement which would lead to a very high quality linkage. In other cases a match may require recoding of keys, resulting in a loss of information from one set of data and a lower quality linkage. An example would be where data in one dataset are coded with a low level of geography, say census blocks, and in the other dataset at a higher level, say a group of blocks. Matching might also require some sort of interpretive coding of data from one dataset. Any such procedure should be documented.
- DDI may need a specific element to contain the project proposal. Currently, *Description* in *FundingInformation* can be used for this purpose; additionally a related *OtherMaterial* can describe further material.
- The *DDI Comparison* element needs additions. The list of target elements needs to expand to include higher level elements like *StudyUnits*, *ControlConstructs*, *Instruments*, *Methodology*, and *Coverage*.
- DDI needs the ability to describe the process (e.g., code) used to subset data (in terms of a selection of cases) for a snapshot. Could *ControlConstruct* be used for this, i.e. inside of *PhysicalDataProduct*? A subset of variables can be already done in *RecordLayout* in *PhysicalDataProduct*.
- DDI may need a machine-actionable method to describe a relationship between a time variable and another variable, for example, a spell.
- DDI should probably have an *InstrumentScheme* element to group *Instruments*.
- It would be desirable for DDI to be able to document and drive and or validate process flow for projects, new samples, quality checks, and event-based data collection. An example of the latter would be a study in which respondents were to receive a survey sometime after some future life event

(e.g., giving birth). Could *Control/Construct* be tweaked for this? Perhaps this could be done by leveraging other standards (e.g., Business Process Management Initiative [BPMN] <http://www.bpmn.org/>).

- DDI should have some way of documenting that a set of data has been destroyed for those projects where destruction of some data is mandated.
- DDI may need to be able to document changes in access rights to data over time and at the variable level. Respondents might revoke previously given rights if asked again. Embargo may not give fine enough control. For fine-grained access control, it is recommended to use other standards like XACML (www.oasis-open.org/committees/xacml/) for (role based) access control and authorization policies. A binding to identifiable DDI objects would be necessary.

Need for DDI Documentation

A DDI Handbook with the core documentation and links to best practice and use case documents would be useful.

CHALLENGES:

What is Longitudinal?

The authors discussed at length just what is meant by a longitudinal study (see Glossary for some standard definitions). A retrospective study might, for example, collect data about people's memories of events over a long time frame. We concluded that if the data were collected in one session, though, this would not be the type of study we wished to consider here. Nevertheless, some of the mentioned best practices can be applied to retrospective studies as well.

Measuring and Recording Time

- Another point of discussion focused on issues related to the measurement of time either at a point or for a spell as tied to some other measurement. At some point does measurement related to a time become invalid? How would this be documented?
- Also discussed was inconsistent measurement of time – a time might be measured relative to Coordinated Universal Time (UTC), as local time without reference to a time zone (or summer adjustment), or simply as an ordinal such as “time 1”. Without associated metadata, replication of the study might be difficult. Representing date and time according to ISO 8601 can solve these issues. Date and time in the metadata expressed in DDI must be specified anyway according to ISO 8601.
- Precision of measurement of time is also important. An example might involve comparing prospective with retrospective measures.

Extended Project-Related Issues

- The authors considered the possibility that with an ongoing study access rights to the data might change over time, including retroactively. A respondent might revoke permission for already collected data to be maintained.
- Similarly what happens when the variable name generation scheme becomes unmaintainable, as for instance at the 11th wave in the series height1, height2... where the wave is represented by one numeric digit? One solution would be to change the structure of variable names and to capture the mapping to the old variables in *Comparison*. The structure (old and new) of naming variables can be documented in a Note related to *VariableScheme*.
- We spent time talking about situations where the obtained universe (sampling frame) for a study or a variable might change. The discussion made the distinction between the intended universe and the obtained universe. An example would be where during the course of a study, investigators discovered that some subset of the intended population was being excluded and then adjusted the procedure to include that subset.
- During the course of a long study DDI itself will change. In other situations investigators may discover that some aspect of their metadata may necessitate an update to DDI or a reconceptualization of how it is used. The addition of biomarker data to ongoing studies is an example. The paper on non-survey data from this workshop (Block et al) will have other examples. Upon consideration, the *Instrument* element was found to be applicable beyond the notion of a survey instrument.

- Versioning and versioning strategy are important topics. The flexibility of the version attribute, which increases in DDI 3.2, means that different investigators may use version in different ways. It will be important to establish a versioning policy and to document it.

Managing the Project

It was determined that studies conducted over a long period will need tools for project management that could, at least in part, be driven by metadata. Some of the issues are:

- Over the course of a long study, the staff involved in gathering and managing the data will likely change. Documenting the provenance of the data will be important.
- Documentation of study evaluation, outcomes, publications, and products using the data will need to be recorded. The Web page for Medical Research Council - MRC e-Val 2009 is an example.
- Linking to external data, e.g., notification of respondent death, will be need to be managed in many cases. Third parties may be involved in cases of deidentified data. Their procedures should be documented. The process for creation and the management of intermediate variables will need to be documented.
- The computation and use of quality assurance metrics (e.g., number of completed observations) will need to be documented.
- Disclosure risk management measures metadata will need to be flexible. Obfuscation technique and the process for generating public disclosure products will need to be recorded. Further exploration into the ability of DDI to capture this information is warranted. Other standards do exist for the management of access rights.
- Documenting reuse of previous data in future data collection is an important practice.

Other Issues

- In a longitudinal study *CollectionEvents* might need to reference multiple *Instruments*.
- There is a need for the ability to apply *Comparison* to high level objects like *StudyUnit* and *DataCollection*. It makes sense to describe commonalities and differences on this level for resource discovery.

SUPPORTING DOCUMENTATION/REFERENCES:

- Amin, Alerk [DDI-users] a question about schemes/versions/references , posting to the DDI-users listserv list <http://www.icpsr.umich.edu/mailman/listinfo/ddi-users> 2010-12-16
- Block, William C., Christian Bilde Andersen, Daniel E. Bontempo, Arofan Gregory, Stan Howald, Douglas Kieweg, and Barry T. Radler. "Documenting a Wider Variety of Data Using the Data Documentation Initiative 3.1: Best Practices, Examples, and Recommendations for Extending the Standard." DDI Working Paper Series, Longitudinal Data Best Practices, Number 1, December 2010. <http://dx.doi.org/10.3886/DDILongitudinal01>
- Data Documentation Initiative. DDI 3.1 Schema and Documentation - Part I – Overview, Part II User Guide, Schema and Field Level Documentation. <http://www.ddialliance.org/specification/ddi3.1#3.1schema>
- Digital Curation Centre. DCC Curation Life Cycle Model. <http://www.dcc.ac.uk/resources/curation-lifecycle-model>
- Edwards, Michelle, Janet Eisenhauer, Jane Fry, Pascal Heus, Kirstine Kolsrud, Meinhard Moschner, Ron Nakao, and Wendy Thomas. "Versioning and Publication." DDI Working Paper Series -- Best Practices, No. 8 (2009-03-22). <http://dx.doi.org/10.3886/DDIBestPractices08>
- Goebel, Jan, and Joachim Wackerow. "New Frontiers: Can Panel Studies Go DDI? First Experiences in Documenting the German Socio-Economic Panel Study With DDI 3.0." Part of session "Prospects for DDI -- What the Evidence and Experience Tell Us," Chair Ron Nakao. Presented at the annual meeting of the International Association of Social Science Information Service and Technology, Montreal, Quebec, May 2007 <http://www.edrs.mcgill.ca/IASSIST2007/presentations/E4%281%29.pdf>
- International Organization for Standardization. ISO/IEC 9834-8:2005 - Information technology -- Open Systems Interconnection -- Procedures for the operation of OSI Registration Authorities: Generation and registration of Universally Unique Identifiers (UUIDs) and their use as ASN.1 Object Identifier components http://www.iso.org/iso/catalogue_detail.htm?csnumber=36775
- Iverson, Jeremy. "Metadata-Driven Survey Design." *IASSIST Quarterly*, Spring/Summer 2009. <http://www.iassistdata.org/iq/metadata-driven-survey-design>
- iShare – Home, INDEPTH Network. <http://www.indepth-ishare.org/>
- Medical Research Council - MRC e-Val 2009. <http://www.mrc.ac.uk/Achievementsimpact/Outputsoutcomes/MRCe-Val2009/index.htm>
- Jääskeläinen, Taina, Meinhard Moschner, and Joachim Wackerow. "Controlled Vocabularies for DDI 3: Enhancing Machine-Actionability." *IASSIST Quarterly*, Spring/Summer 2009. http://iassistdata.org/downloads/iqvol3312wackerow_0.pdf
- OECD Glossary of Statistical Terms. <http://stats.oecd.org/glossary/>
- SDMX Content-Oriented Guidelines Annex 4: Metadata Common Vocabulary 2009. http://sdmx.org/wp-content/uploads/2009/01/04_sdmx_cog_annex_4_mcv_2009.pdf
- Wackerow, Joachim. "Comparison | DDI - Data Documentation Initiative" (Example showing the grouping approach for comparable variables, description of derived variables as well as the relationship of waves

and the household / person relationship) DDI3.0 example file (valid according to the public review version of DDI 3.0). <http://www.ddialliance.org/specification/proof-of-concept> (see SOEP Panel Study)

APPENDIX A

GLOSSARY

Administrative data	Data collected for the administration of government (or other) programs. Examples include: <ul style="list-style-type: none"> ▪ Economic data ▪ Educational achievement in public schools ▪ Hospital admissions/discharges/outcomes ▪ Income/sales/property tax records (both personal and business) ▪ Immigration applications/approvals/naturalization records ▪ Social Security records ▪ Unemployment Insurance claims/records ▪ Voting records ▪ Workers compensation (for on-the-job injuries)
Biomarker	<p>The official NIH definition of a biomarker is: “a characteristic that is objectively measured and evaluated as an indicator of normal biologic processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention.”</p> <p>Ref: Biomarkers Definitions Working Group: “Biomarkers and Surrogate Endpoints: Preferred Definitions and Conceptual Framework.” CLIN PHARMACOL THER 2001;69:89-95.</p> <p>http://www.everythingbio.com/glos/definition.php?ID=3716</p>
Cohort/Event-based	Data collected over time about a group of individuals that are connected in some way or have shared some significant experience within a given period. Examples: birth, disease, education, employment, family formation, participation in an event.
Concordance	Tool or table indicating the presence of the same variable or question over waves of a study.
Continuous panel	Reports from a panel collected on a regular basis.
Continuous time series	Phenomena measured at every instant of time. Examples: lie detectors, electrocardiograms, etc.
Cross-sectional	Data about a population obtained only once.
Cross-sectional ad-hoc followup	Data collected at one point in time to complete information collected in a previous cross-sectional study; the decision to collect follow-up data is not included in the study design.

Data harmonization	Data harmonization is the process of bringing variable-level information into alignment to express comparability. This is often done through mapping across various elements of the variables, including variable name, label, categories, codes, etc.
Data life cycle	The whole course of existence of a set of data, from initial conception to ultimate disposal.
DDI	The Data Documentation Initiative (http://www.ddialliance.org/). Also that organization's metadata specification for the social and behavioral sciences.
Digital Object Identifier (DOI)	A character string used to uniquely identify an electronic document or other object. Metadata about the object is stored in association with the DOI name and this metadata may include a location, such as a URL, where the object can be found. The DOI for a document or dataset is permanent, whereas its location and other metadata may change. Referring to an online document by its DOI provides more stable linking than simply referring to it by its URL, because if its URL changes, the publisher need only update the metadata for the DOI to link to the new URL.
Discrete time series	Measurements taken at (usually regularly) spaced intervals.
DSS / HDSS	Health and Demographic Surveillance Systems (HDSS) for longitudinal monitoring of small-area populations by continuous recording of vital events have been set up in many developing countries. HDSS's are based on a data gathering method comprising an initial census of the resident population, followed by multi-round surveys covering all inhabitants of the area. They thus, provide a geographical and temporal observation window on a locally circumscribed population defined using certain rules of residence. Individuals' life events during their period(s) of residence in the survey area are recorded on an individual basis (the minimum data being births, deaths and migration), but sometimes per household or per residential unit. Examples: macroeconomics (weekly share prices, monthly profits, sales); meteorology (daily rainfall, hourly temperature); measurements of individuals (blood pressure, weight, height); sociology (crime figures, employment figures), etc.
Grouping	A DDI mechanism to clearly document the repurposing of aspects of the initial study and the relationships that exists between each of the component studies in the group. The typical use case involves a series or collection of studies which are related in some way or a group of studies which are being compared. A Group can be comprised of StudyUnits and SubGroups. A standard set of attributes describes the following dimensions for grouping: Time, Instrument, Panel, Geography, Datasets, Language.

Instrument	A specific instrument or tool used to collect data. For survey data, the instrument has traditionally been seen as a questionnaire, but devices used to collect biomedical information, e.g., fMRI scanning devices, can also be viewed as instruments.
Interval panel	Measurements taken only when information is needed.
Longitudinal	Data collected repeatedly over time to study change in a population.
Panel	Data collected over time from, or about, the same sample of respondents.
Published	The DDI attribute <code>isPublished</code> is set to true when the metadata are made available outside of the group of original developers. Published metadata must be versioned.
Register data	Data collected and maintained on individuals and businesses to track vital statistics and other information.
Resource package	A means of packaging any maintainable set of DDI metadata for referencing as part of a study unit or group. A resource package structures materials for publication that are intended to be reused by multiple studies, projects, or communities of users. A resource package uses the group module with an alternative top-level element called Resource Package that is used to describe maintainable modules or schemes that may be used by multiple study units outside of a group.
Retrospective study	A study in which data are collected from recollections of past events.
Surveillance study	A study in which data are collected by systematic observation.
Time series	Data collected repeatedly over time to study change in observations. These are typically “objective” measurements of phenomena that can be observed externally, as opposed to attitudes/opinions or feelings. Examples may include economic/financial indicators, natural/meteorological phenomena, vital statistics, etc.
Trend/Repeated cross-section	The study of different samples/different groups of people from the same population at several points in time, using the same set of questions/variables. Conclusions are drawn for the population. Examples: public opinion polls, elections studies, etc.
Trials / Interventions	A study involving some sort of experimental action usually in comparison to some control condition.
Versioned	Metadata for which any changes will require an update of the version attribute of the metadata.
Wave	One of a sequence of repeated stages of a study.

APPENDIX B

ACKNOWLEDGMENTS

The paper is one of several papers that are the outcome of a workshop held at Schloss Dagstuhl - Leibniz Center for Informatics in Wadern, Germany, on October 18-22, 2010. The series was edited by Stefan Kramer, Larry Hoyle, and Mary Vardigan.

Workshop Title:

The Data Documentation Initiative (DDI) Standard : Managing Metadata for Longitudinal Data — Best Practices

Link: <http://www.dagstuhl.de/10422>

Organizers:

Arofan Gregory (Open Data Foundation, Tucson, Arizona, USA)

Mary Vardigan (Inter-university Consortium for Political and Social Research [ICPSR], University of Michigan, USA)

Joachim Wackerow (GESIS, Leibniz Institute for the Social Sciences, Germany)

Participants in the workshop:

- Christian Bilde Andersen, Danish Data Archive (DDA)
- Randy Banks, Institute for Social and Economic Research (ISER), University of Essex
- Bill Block, Cornell Institute for Social and Economic Research (CISER), Cornell University
- Daniel Bontempo, Life Span Institute, University of Kansas
- Fortunato Castillo, MRC Centre of Epidemiology for Child Health, Institute of Child Health, University College London
- Vicky (Huey-Chi) Chang, Wisconsin Longitudinal Study, University of Wisconsin-Madison
- Benjamin Clark, London School of Hygiene and Tropical Medicine, Tazama Project, Tanzania
- Sue Ellen Hansen, Institute for Social Research, Survey Research Operations, University of Michigan
- Stan Howald, Wisconsin Longitudinal Study, University of Wisconsin-Madison
- Larry Hoyle, Institute for Policy and Social Research, University of Kansas
- Jeremy Iverson, Algenta Technologies
- Uwe Jensen, GESIS - Leibniz Institute for the Social Sciences
- Douglas Kieweg, Center for Biobehavioral Neurosciences in Communication Disorders (BNCD), University of Kansas
- Neeraj Kumar Kashyap, Vadu Rural Health Program, KEM Hospital Research Centre, INDEPTH Network
- Stefan Kramer, Cornell Institute for Social and Economic Research (CISER), Cornell University
- Hilde Orten, Norwegian Social Science Data Archive (NSD)
- Denise Perpich, Language Acquisition Studies Lab, University of Kansas
- Barry Radler, Institute on Aging, University of Wisconsin-Madison
- Ingo Sieber, German Institute for Economic Research (DIW) Berlin, Socio-Economic Panel Study (SOEP)
- Johanna Vompras, University Bielefeld Library, Germany
- Knut Wenzig, National Educational Panel Study (NEPS), University of Bamberg
- Wolfgang Zenk-Möltgen, GESIS - Leibniz Institute for the Social Sciences

APPENDIX C

Copyright © DDI Alliance 2011, *All Rights Reserved*

<http://www.ddialliance.org/>

Content of this document is licensed under a Creative Commons License:
Attribution-Noncommercial-Share Alike 3.0 United States

This is a human-readable summary of the Legal Code (the full license).

<http://creativecommons.org/licenses/by-nc-sa/3.0/us/>

You are free:

- to Share - to copy, distribute, display, and perform the work
- to Remix - to make derivative works

Under the following conditions:

- Attribution. You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).
- Noncommercial. You may not use this work for commercial purposes.
- Share Alike. If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar license to this one. For any reuse or distribution, you must make clear to others the license terms of this work. The best way to do this is with a link to this Web page.
- Any of the above conditions can be waived if you get permission from the copyright holder.
- Apart from the remix rights granted under this license, nothing in this license impairs or restricts the author's moral rights.

Disclaimer

The Commons Deed is not a license. It is simply a handy reference for understanding the Legal Code (the full license) — it is a human-readable expression of some of its key terms. Think of it as the user-friendly interface to the Legal Code beneath. This Deed itself has no legal value, and its contents do not appear in the actual license.

Creative Commons is not a law firm and does not provide legal services. Distributing of, displaying of, or linking to this Commons Deed does not create an attorney-client relationship. Your fair use and other rights are in no way affected by the above.

Legal Code:

<http://creativecommons.org/licenses/by-nc-sa/3.0/us/legalcode>