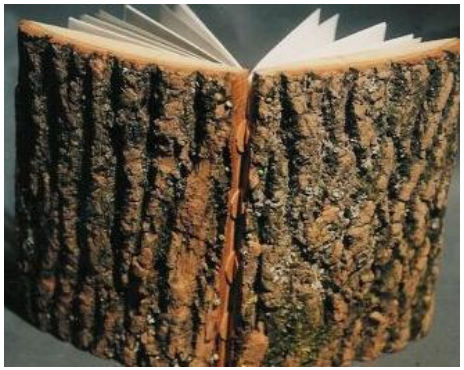
İstatistiksel Veri ve Üst veri Modelleri

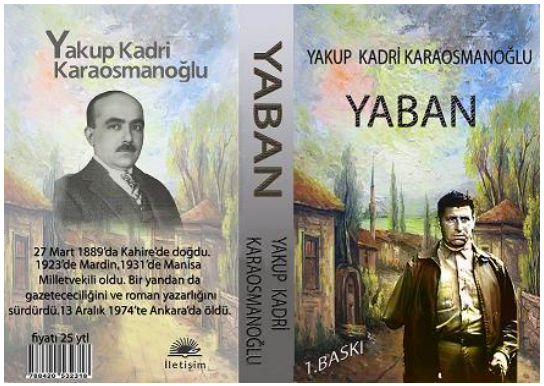
Akademik Bilişim konferansı 2015
Eskişehir
Akın ÖZTÜRK
Tuğba TUĞCU

Bir Standartlaştırma Aracı Olarak üst veri

üst veri Nedir?

üst veri → Veri hakkındaki her türlü veri





74724269 → 31 Aralık 2011 tarihi itibariyle Türkiye'nin toplam nüfusu

- üst veri, veriyi tanımlayan ve anlamlandıran **üst bilgidir**.
- üst veri, veriyi **bilgiye dönüştüren** şeydir.
- üst veri, **veri ile bilgi arasındaki boşluğu kapatmak** için kullanılan bir teknoloji veya yaklaşımdır.
- üst veri, en genel tanımıyla verinin anlamını, kalitesini, kaynağını, formatını ve bunun gibi değişkenleri tanımlayan veridir.

Neden üst veri?

Etkin bir **üst veri yaklaşımı**, **veri tufanı** karşısında sağlam bir **bariyerdir**.

Standartlaşma

- **Ortak Süreçler**
- **Ortak Kavramlar**
- **Ortak Değişkenler**
- **Ortak …**

Kurumsal Hafıza

- **Dokümantasyon**
- **Yeniden Kullanım**

Birlikte Çalışılabilirlik

Ulusal İstatistik Sistemi içindeki koordinasyon ve entegrasyon

Kalite

üst veri Kavramları

İstatistiksel Veri Türleri

Mikro

- **Ham Veri:** Cevaplayıcıdan alınan veridir. Hiçbir işlem görmemiştir.
- **Mikro Veri:** Ham verinin temizlenmiş ve kodlanmış halidir.

Makro

- **Makro Veri:** Mikro verinin düzenlenmiş ve birleştirilmiş halidir.
- **Yayın (Dağıtım) Verisi:** Makro verinin analiz edilmiş ve görselleştirilmiş halidir.

üst veri Türleri:

- **Yapısal üst veri:** Verinin yapısını tanımlamak için kullanılan üst veridir.

  Değişken isimleri, Kod listeleri, Veri türleri, Veri seti tanımları, Ölçüm birimi, Değer aralığı vb.

- **Referans üst veri:** İstatistiksel verinin içeriğini ve kalitesini tanımlayan üst veridir.

  İstatistiksel verinin bağlamı üzerinde açıklayıcı metinler, veri toplama yöntemleri, veri işleme yöntemleri, kalite ve dağıtım göstergeleri

  - **Kavramsal üst veri:** Kullanılan kavramları içerir.
  - **Yöntemsel üst veri:** İstatistiksel verinin oluşturulması için kullanılan yöntemlerin belirlendiği üst veridir.

    Örnekleme yöntemleri, Veri toplama yöntemleri, veri düzenleme süreçleri, tahmin yöntemleri vb.
  - **Kalite üst verisi:** Oluşturulan istatistiğin kalite göstergelerinin tanımlandığı üst veridir.

    Güncellik, doğruluk, tamlık vb.

Uluslararası üst veri Standartları

- Statistical Data and Metadata Exchange (SDMX): Makro, toplulaştırılmış veri modellemek için kullanılmaktadır.
- Data Documentation Initiative (DDI): Micro veri ve dokümanların tüm ayrıntılara kadar ifade etmede kullanılmaktadır.
- ISO 11179: Semantik modelleme ve üst verinin kalite bileşenlerinin(kavramların, tanımların vb.) tanımlanması, veri gösterimi kullanılmaktadır.
- ISO 19115: Coğrafi bilgi sistemlerinin veri modellerinde kullanılmaktadır.
- Dublin Core: Görsel medya, dosyalar için üst veri modellenmesinde kullanılmaktadır.

DDI (Data Documentation Initiative)

- DDI ekonomi, sosyal ve davranış bilimlerinde kullanılan ve oluşturulan verilerin tanımlanması ve dokümante edilmesi için geliştirilmiş bir XML standardıdır.

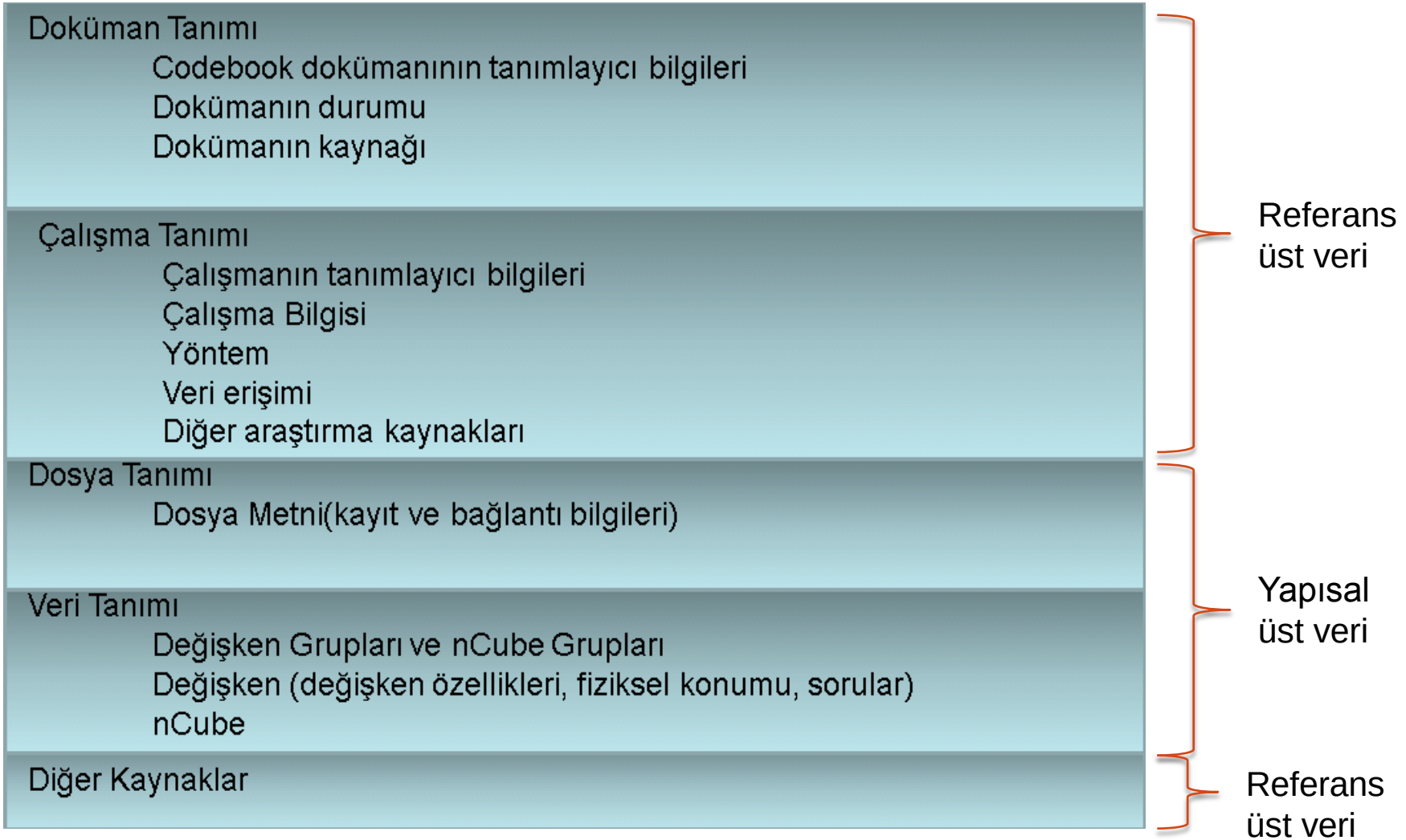
DDI-Codebook

1. Nesil DDI 1.0…2.1 (2000-2008)
  - Arşivleme odaklı.
  - Bir çalışmanın/veri setinin/ anketin sunumu.

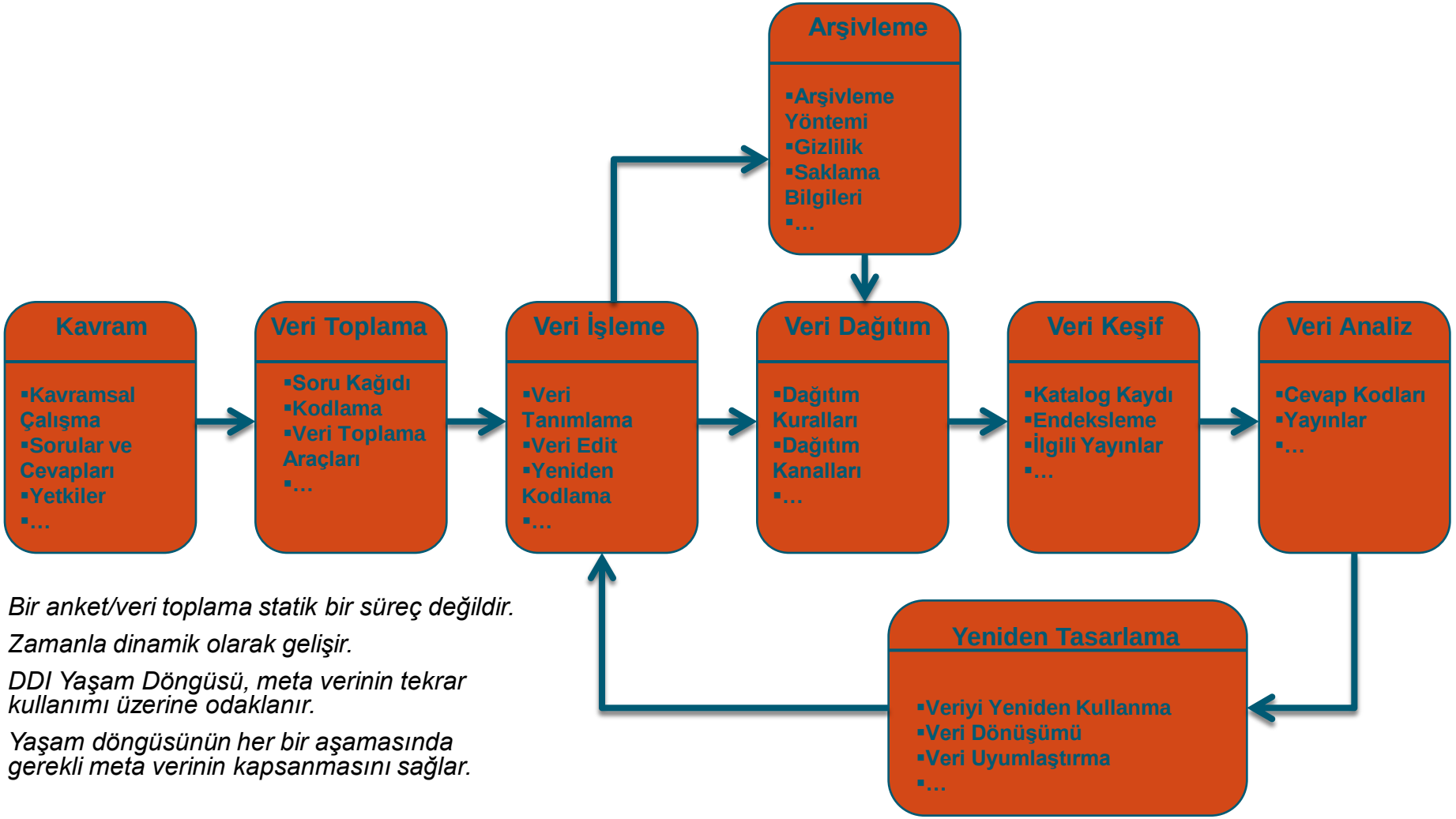
DDI-Lifecycle

2. Nesil DDI 3.0 -3.1 (2008)
  - Yaşam döngüsü odaklı
  - Tek anket/çalışma/veri seti kavramı ötesinde
  - Örnekleme, Anket tasarımı, veri kalitesinin dokümantasyonu üzerinde çalışmalar

DDI-Codebook (Kod Rehberi) Yapısı



DDI-Lifecycle (Yaşam Döngüsü) Yapısı



Bir anket/veri toplama statik bir süreç değildir.

Zamanla dinamik olarak gelişir.

DDI Yaşam Döngüsü, meta verinin tekrar kullanımı üzerine odaklanır.

Yaşam döngüsünün her bir aşamasında gerekli meta verinin kapsanmasını sağlar.

DDI-Lifecycle meta veri tipleri

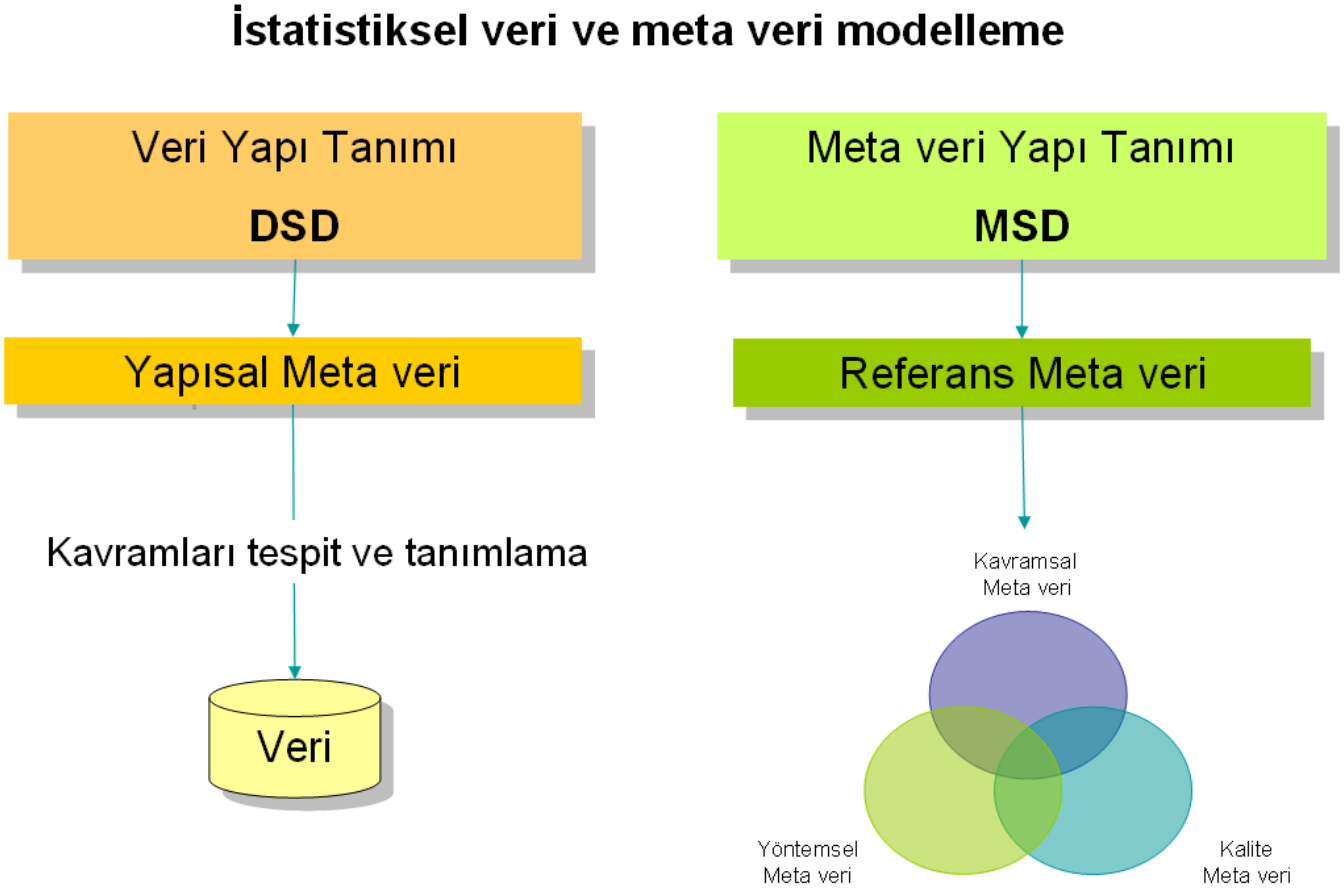
- **Kavramlar** / Concepts (“terms”)
  - *Kavramlar bir araştırmada kullanılan bütün terimleri ifade eder. Bir araştırmada her bir kavramın adı, etiketi ve açıklaması gerekli alanlardır. Aynı kavram farklı araştırmalarda kullanılabilir. Bunun için bir kavram bankası oluşturulmalıdır.*
- **Araştırmalar**/ Studies (“surveys”, “collections”, “samples”, “censuses” etc)
  - *Araştırma bir veri toplama sürecini ve süreç sonunda ortaya çıkan sonuçları kapsar.*
  - *DDI modelinde bir araştırma şu meta verileri içerir.*
    - ***Abstract(Özet) :*** *Araştırmanın kapsam ve içeriğinin özeti.*
    - ***Citation(Atıf):*** *Araştırmanın başlığı, altbaşlığı, oluşturan, yayınlayan bilgileri gibi atıf bilgilerini içerir.*
    - ***Universe (Evren) :*** *Araştırmaya konu olan kişi ve nesnelerin açıklamasını içerir.*
    - ***SeriesStatement:*** *Araştırmanın bağlı olduğu araştırma serisini içerir.*
    - ***FundingInformation(Finansal Bilgiler):*** *Araştırmanın finansal bilgilerini içerir.*
    - ***Purpose (Amaç):*** *Araştırmanın amacını içerir.*
    - ***Coverage(Kapsam):*** *Araştırmanın konuya, zamana ve mekana ilişkin kapsamını içerir.*
    - ***AnalysisUnit (Analiz Birimi):*** *Araştırmadaki analiz birimlerini içerir.*
    - ***KindOfData (Veri Tipi):*** *Araştırmada derlenen verinin tipini gösterir. Örnek: anket verisi, sayım verisi, idari veri, ölçüm verisi, demografik veri vb.*
    - ***ConceptualComponent (Kavramsal Bileşen):*** *Araştırmanın kullandığı kavramlar ve evrenler ilgili meta verileri içeren bölümdür.*
    - ***DataCollection (Veri Toplama):*** *Araştırmanın veri toplama süreciyle ilgili tüm meta verileri içerir. Bu bölümde veri kaynaklarını belirlemede kullanılan metodolojiler, soru yapıları ve akışları yer alır.*
    - ***LogicalProduct (Mantıksal Ürün):*** *Araştırmanın değişkenlerinin, kategorilerinin, kod listelerinin belirtildiği bölümdür.*
- **Anket Araçları/ Survey instruments (“questionnaire”, “form”)**
  - *Anket içeriğinin ve yapısının belirtildiği bölümdür.*
- **Sorular / Questions (“observations”)**
- **Cevaplar / Responses**
- **Değişkenler / Variables (“data elements”, “columns”)**
- **Kod Listeleri ve Kategoriler / Codes & categories (“classifications”, “codelists”)**
- **Evren / Universes (“populations”, “samples”)**
- **Veri Setleri / Data files (“data sets”, “databases”)**

SDMX (Statistical Data and Metadata Exchange)

- SDMX, istatistiksel veri ve üst verinin, uluslararası kuruluşlar ve onların üye ülkeleri arasında karşılıklı değişimi ve paylaşımı için daha etkin yöntemler geliştirilmesi ve kullanılmasını sağlayan bir veri değişim standardıdır.
- SDMX girişimi 2001 yılında uluslararası düzeyde istatistik üzerine çalışan yedi kuruluş tarafından başlatılmıştır.

Birleşmiş Milletler İstatistik Bölümü (UNSD)	**Uluslararası Para Fonu (IMF)**
Dünya Bankası (WB)	**Ekonomik İşbirliği ve Kalkınma Örgütü (OECD)**
Uluslararası Ödemeler Bankası BIS (BIS)	**Avrupa Birliği İstatistik Ofisi (Eurostat)**
Avrupa Birliği Merkez Bankası (ECB)	

SDMX Nedir?


İstatistiksel veri ve meta veri modelleme
Veri Yapı Tanımı
DSD
Yapısal Meta veri
Kavramları tespit ve tanımlama
Veri
Meta veri Yapı Tanımı
MSD
Referans Meta veri
Kavramsal
Meta veri
Yöntemsel
Meta veri
Kalite
Meta veri

SDMX tanımlar

- Boyutlar(Dimentions): Taşınacak veri tablo şeklide gösterilirken, ölçülen veriyi tanımlayan ve betimleyen değişkenlerdir. Bu bilgi gönderilmediği zaman değerlerin istatistiksel anlamında değişiklik olması beklenir Örneğin tablonun satır ve sütunlarını.
- Özellikler(Attributes): Taşınacak veri tablo şeklide gösterilirken, ölçülen veriyi sadece betimleyen değişkenlerdir. Bu bilgi gönderilmediği zaman değerlerin istatistiksel anlamında değişiklik olmaması beklenir.
- Değerler(Measures): Taşınacak veri tablo şeklide gösterilirken, ölçüm sonuçlarıdır.

SDMX Bileşenleri-1

- İstatistiksel veriyi, üst veriyi ve veri alışverişi süreçlerini modellemeyi sağlayan **bilgi modeli**
  - Belirli bir istatistiki alan için veri (ve ilgili üst veri), bir **"Veri Yapısı Tanımlaması" na (Data Structure Definition-DSD)** göre yapılandırılmıştır. DSD belirli bir istatistiki veri akışının yapısını, **boyutların** bir listesi (örneğin: ülke, değişken/başlık, yıl), **özelliklerin** bir listesi (örneğin, ölçüm birimi) ve onların ilişkili **kod listelerinden** yararlanarak tanımlar. **Özellikler** tek bir değer, zaman serisi veya bir zaman serisi grubu hakkında üst verilerdir.

BİRİM ZAMAN TURİZM_KONU SIKLIK

Number of establishments, bedrooms and bedplaces - national - annual data

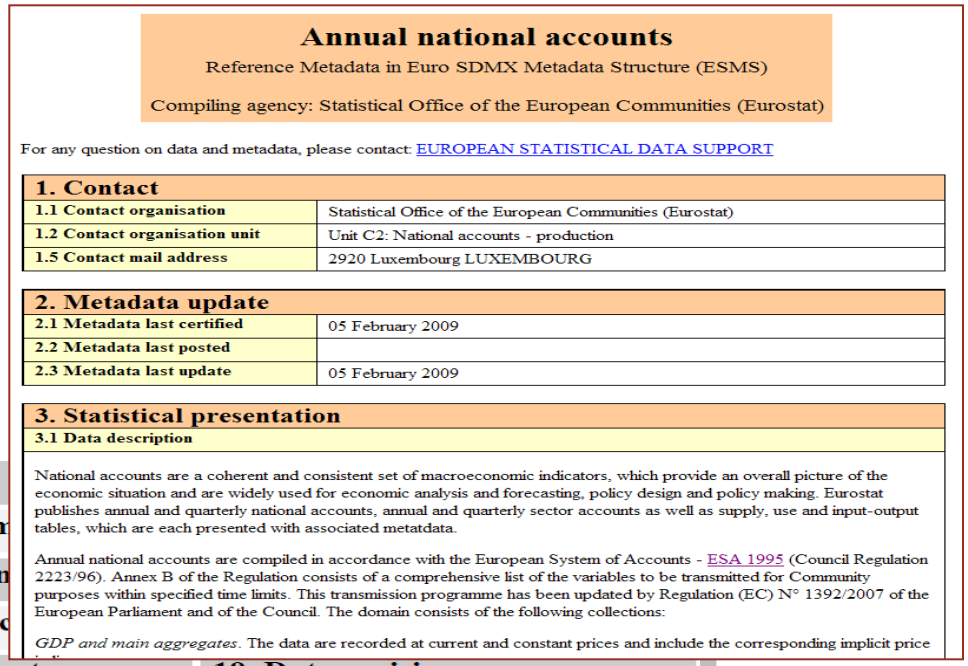
Topic	A100			B010			B020		
Country/Time	2005	2006	2007	2005	2006	2007	2005	2006	2007
AT	14267	14051	14204	538	542	540	3225	3329	3388
ES	17607	18304	17827	1250	1216	1220	4552	4524	4843
FR	18689	18361	18135	8174	8138	8052	2329	2325	2406
IT	33527	33768	34058	2411	2510	2587	68385	68376	61810

ÜLKE ÖLÇÜLEN_DEĞER

BOYUTLAR ÖZELLİKLER DEĞER

SDMX Bileşenleri-2

- Referans üst veriyi tanımlamak için **“üst veri Yapısı Tanımlaması” na (Metadata Structure Definition-MSD) kullanılır.**



Annual national accounts

Reference Metadata in Euro SDMX Metadata Structure (ESMS)

Compiling agency: Statistical Office of the European Communities (Eurostat)

For any question on data and metadata, please contact: EUROPEAN STATISTICAL DATA SUPPORT

1. Contact	
1.1 Contact organisation	Statistical Office of the European Communities (Eurostat)
1.2 Contact organisation unit	Unit C2: National accounts - production
1.5 Contact mail address	2920 Luxembourg LUXEMBOURG

2. Metadata update	
2.1 Metadata last certified	05 February 2009
2.2 Metadata last posted	
2.3 Metadata last update	05 February 2009

3. Statistical presentation

3.1 Data description

National accounts are a coherent and consistent set of macroeconomic indicators, which provide an overall picture of the economic situation and are widely used for economic analysis and forecasting, policy design and policy making. Eurostat publishes annual and quarterly national accounts, annual and quarterly sector accounts as well as supply, use and input-output tables, which are each presented with associated metatdata.

Annual national accounts are compiled in accordance with the European System of Accounts - ESA 1995 (Council Regulation 2223/96). Annex B of the Regulation consists of a comprehensive list of the variables to be transmitted for Community purposes within specified time limits. This transmission programme has been updated by Regulation (EC) N° 1392/2007 of the European Parliament and of the Council. The domain consists of the following collections:

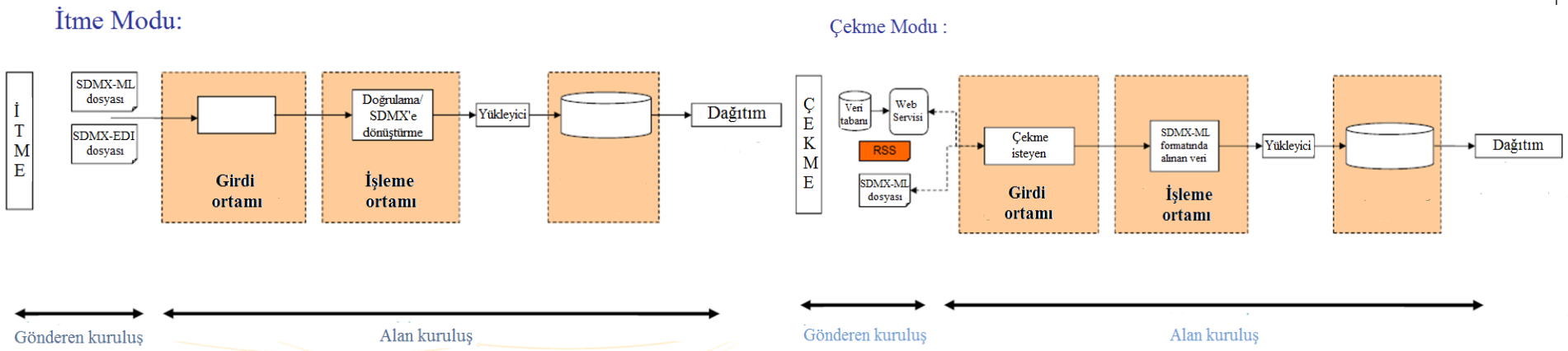
GDP and main aggregates. The data are recorded at current and constant prices and include the corresponding implicit price

üst veri temel kavramları

1. Contact	8. Release policy	
2. Metadata update	9. Frequency of dissen	
3. Statistical presentation	10. Dissemination forn	
4. Unit of measure	11. Accessibility of doc	
5. Reference period	12. Quality management	19. Data revision
6. Institutional mandate	13. Relevance	20. Statistical processing
7. Confidentiality	14. Accuracy and reliability	21 Comment

SDMX Bileşenleri-3

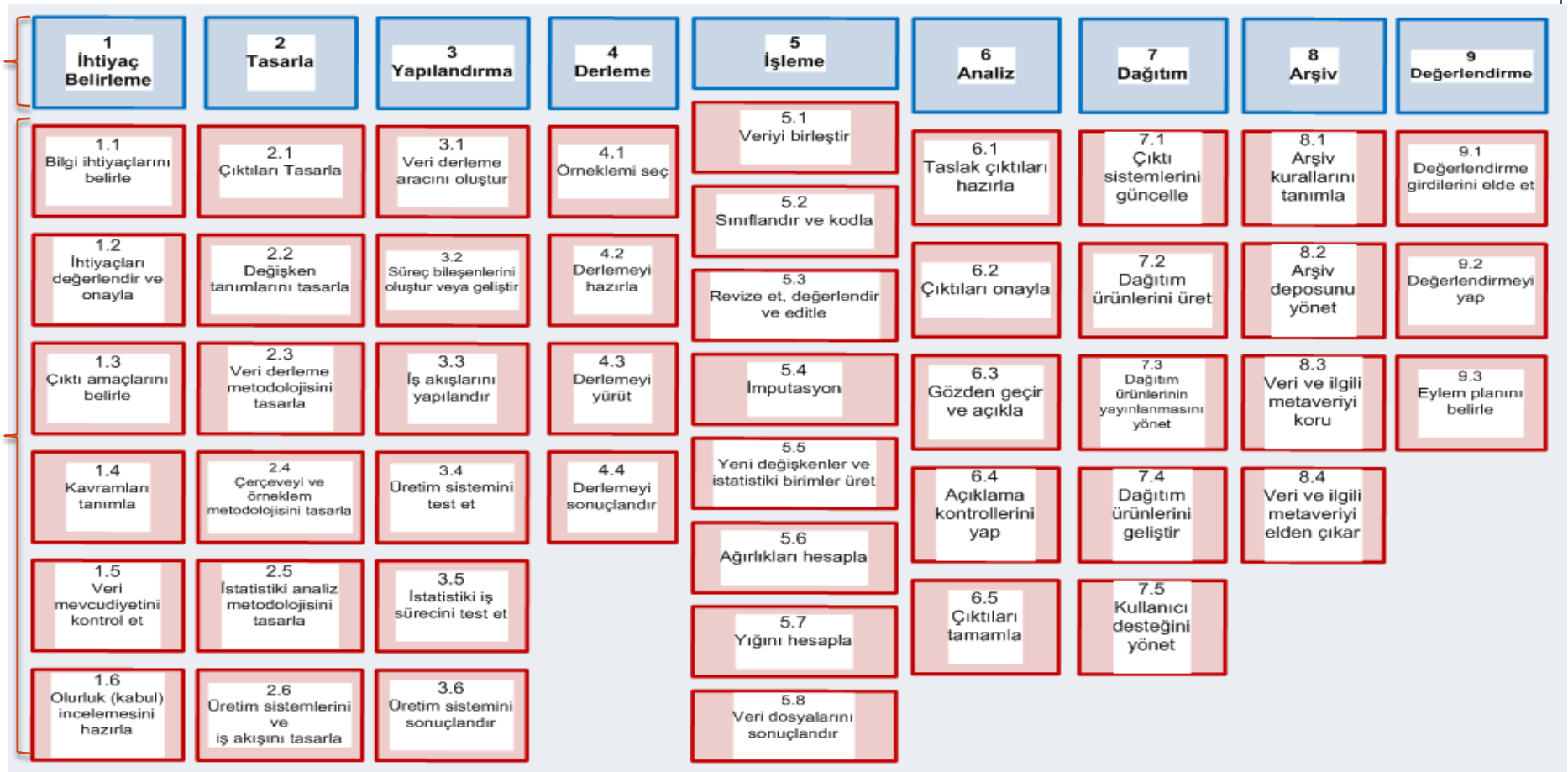
- Veri ve üst veri değişimi için standart formatlar (EDI, XML)
- İçerik Odaklı Kılavuzlar (verinin kategorilenmesi ve tanımlanması için **tavsiyeler**)
  - İçerik odaklı kılavuzlar verinin kategorilenmesi ve tanımlanması için **tavsiyelerdir**.
    - **Tüm Alanlarda Geçerli Kavramlar**
    - **Tüm Alanlarda Geçerli Kod Listeleri**
    - **İstatistiki Konu Alanları**
    - **üst veri Ortak Kelime Haznesi**
- Veri değişimi ve paylaşımı için IT mimarisi



GSBPM(Generic Statistical Business Process Model)

- UN/ECE (Birleşmiş Milletler Avrupa Ekonomik Komisyonu) bünyesindeki METIS (İstatistiksel üst veri) grubu, istatistik üretim süreci için jenerik bir referans modeli oluşturmuştur.
- GSBPM Resmi istatistik üreticileri tarafından istatistik üretimi için yapılan tüm faaliyetleri içerir.
- Veri kaynağından bağımsız olarak modellenmiştir. Anketler, sayımlar, idari kayıtlar, karma kaynaklar, istatistiki kayıtların bakımı, vb. kaynaklardan istatistik üretimi için kullanılabilir.
- Bir istatistiki süreç, evre ve alt süreçlerden oluşur.
- Alt süreçler, sıkı bir sırada takip edilmek zorunda değildir. Fazlar arasında farklı alt süreçler belirli sıralarla yürüyebilir.

GSBPM


Evre
Alt Süreçler
1 İhtiyaç Belirleme
2 Tasarla
3 Yapılandırma
4 Derleme
5 İşleme
6 Analiz
7 Dağıtım
8 Arşiv
9 Değerlendirme
1.1 Bilgi ihtiyaçlarını belirle
1.2 İhtiyaçları değerlendir ve onayla
1.3 Çıktı amaçlarını belirle
1.4 Kavramları tanımla
1.5 Veri mevcudiyetini kontrol et
1.6 Olurluk (kabul) incelemesini hazırla
2.1 Çıktıları Tasarla
2.2 Değişken tanımlarını tasarla
2.3 Veri derleme metodolojisini tasarla
2.4 Çerçeveyi ve örneklem metodolojisini tasarla
2.5 İstatistiki analiz metodolojisini tasarla
2.6 Üretim sistemlerini ve iş akışını tasarla
3.1 Veri derleme aracını oluştur
3.2 Süreç bileşenlerini oluştur veya geliştir
3.3 İş akışlarını yapılandır
3.4 Üretim sistemini test et
3.5 İstatistiki iş sürecini test et
3.6 Üretim sistemini sonuçlandır
4.1 Örneklemi seç
4.2 Derlemeyi hazırla
4.3 Derlemeyi yürüt
4.4 Derlemeyi sonuçlandır
5.1 Veriyi birleştir
5.2 Sınıflandır ve kodla
5.3 Revize et, değerlendir ve editle
5.4 İmputasyon
5.5 Yeni değişkenler ve istatistiki birimler üret
5.6 Ağırlıkları hesapla
5.7 Yığını hesapla
5.8 Veri dosyalarını sonuçlandır
6.1 Taslak çıktıları hazırla
6.2 Çıktıları onayla
6.3 Gözden geçir ve açıkla
6.4 Açıklama kontrollerini yap
6.5 Çıktıları tamamla
7.1 Çıktı sistemlerini güncelle
7.2 Dağıtım ürünlerini üret
7.3 Dağıtım ürünlerinin yayınlanmasını yönet
7.4 Dağıtım ürünlerini geliştir
7.5 Kullanıcı desteğini yönet
8.1 Arşiv kurallarını tanımla
8.2 Arşiv deposunu yönet
8.3 Veri ve ilgili metaveriyi koru
8.4 Veri ve ilgili metaveriyi elden çıkar
9.1 Değerlendirme girdilerini elde et
9.2 Değerlendirmeyi yap
9.3 Eylem planını belirle

Sonuç

- Herhangi bir verinin insanlar ve makineler tarafından anlaşılması, kullanılması ve değişimi için belli standartlara uyması gerekmektedir.
- Veri ve üst veri modellemesine bir çok alanda ihtiyaç vardır.
  - Anket geliştirmede,
  - anketlerin üst bilgisi üretiminde,
  - bilgilerin yayınlanmasında,
  - kurumlar arasında veri transferinde,
  - idari kayıtların değişiminde altlık olarak,
  - veri ambarı ve veri madenciliği (veri deseni) veri modellerine ihtiyaç vardır.

Teşekkürler

- akin2100@gmail.com