DataForge

IASSIST 2013
The <Meta>Data Dance
The `<Meta/>` Data Dance
The `<Meta>`Data Dance
Tools to facilitate statistical/scientific data management
- Complement existing tools (not replace)
- Facilitate adoption of standards and promote best practices
- Alleviate the need to master complexities of metadata standards, technologies

Vision: web based services (cloud users), stand alone desktop / server products (agencies/enclaves), libraries (developers)

Broad Target audience: Data Producers / Manager / Researchers / Users

DataForge Today (lean model):
- Simple desktop based command line tools
  - SledgeHammer - Data/metadata transformation
    - LavaCore - Java library
  - Caelum - Reporting / Publication

DataForge Tomorrow:
- DataForge Online, Web Services, Desktop w/UI, Java Library
Data Liberation
- Unlock from proprietary formats, turn into open data
- Produce metadata + convert data into standard ASCII formats

Open <Meta>Data Integration
- Generating scripts for loading into stats packages, databases, big data engines, cloud
- Use with open systems / environments
Challenges

- Read proprietary formats
- Data integrity
- ASCII optimization
- Cross package formatting issues
- SQL Database limits (columns, indexes, views)
- Performance (Timings, size, memory usage)
- ...
Key Features

- Input: SPSS, Stata, ASCII+SAS syntax, ASCII+DDI-C / DDI-L, ASCII+SSS 1.1/2.0, ASCII+StatTransfer
- Data Out: Fixed w/optimization, CSV, Delimited
- Metadata Out: DDI-C 1.2.2/Nesstar/2.1/2.5, DDI-L 3.1-RP/3.1-SU/Colectica, Triple-S 1.1/2.0
- Summary Stats: Min / Max /Valid/Invalid /Mean, StdDev /Variance / Frequencies, Weighted *, Save in CSV*.
- Scripts Generators: SAS, SPSS, Stata, R, Mysql, Oracle*, MS-SQL*, HSQL*, Google BigQuery*, HP/Vertica* + dimension/lookup tables, indexes/foreign keys*
Editions

Freeware

- Anonymous: 100 variables / 1,000 obs.
- Registered: 500 variables / 5000 obs. (or contact us)

Pro/Enterprise:

- 4Q2013
- No limits, all features, disable usage report
- pricing/licensing t.b.d.
- In private beta (contact us for info)

LavaCore:

- Currently internal only, planning for OEM
Future Features (demand driven)
- Desktop UI
- Support additional input/output formats (SAS, cloud)
- Support missing value
- Additional ID generators
- Compute subsets
- Synthetic data generators
- Disclosure control
- ???

DataForge Online (2014)
- Free and pay as you go services

DataForge Services (available today)
- Community support in Google groups
- MTNA: tech support, custom developments
- IDMS: data/metadata management / processing
MTNA Project Stories

- Generated DDI for American National Election Studies (ANES)
- Loaded US General Social Survey (1972-2010) surveys into Big Query (5K+ variables, 55K records, 2.2Gb, query in seconds)
- Generated MS-SQL and loaded data for KUSP Translating Research in Elder Care (TREC) surveys (U. Alberta, Canada)
- Integration in NSF / NORC SED Manager (LavaCore)
- Integration in Canada RDC Dataset Builder (LavaCore)

...
Caelum: A constellation in the southern sky, “chisel” in latin. Formerly known as Caelum Scalptorium, "the engraver's chisel”.

Rationale: I have DDI, then what…. Not everyone is an XML/technology expert.

Use cases: Produce HTML/PDF reports, quality assurance, XML conversion, etc..

How does it work: Command line tool that wraps standards XML and other transformation technologies. Can be used with any XML

Comes with demo transforms with objective to foster community contributions

Download at http://www.openmetadata.org/dataforge/caelum
http://www.openmetadata.org/dataforge

DEMO