Standardisation Challenges in Data Management Lifecycle

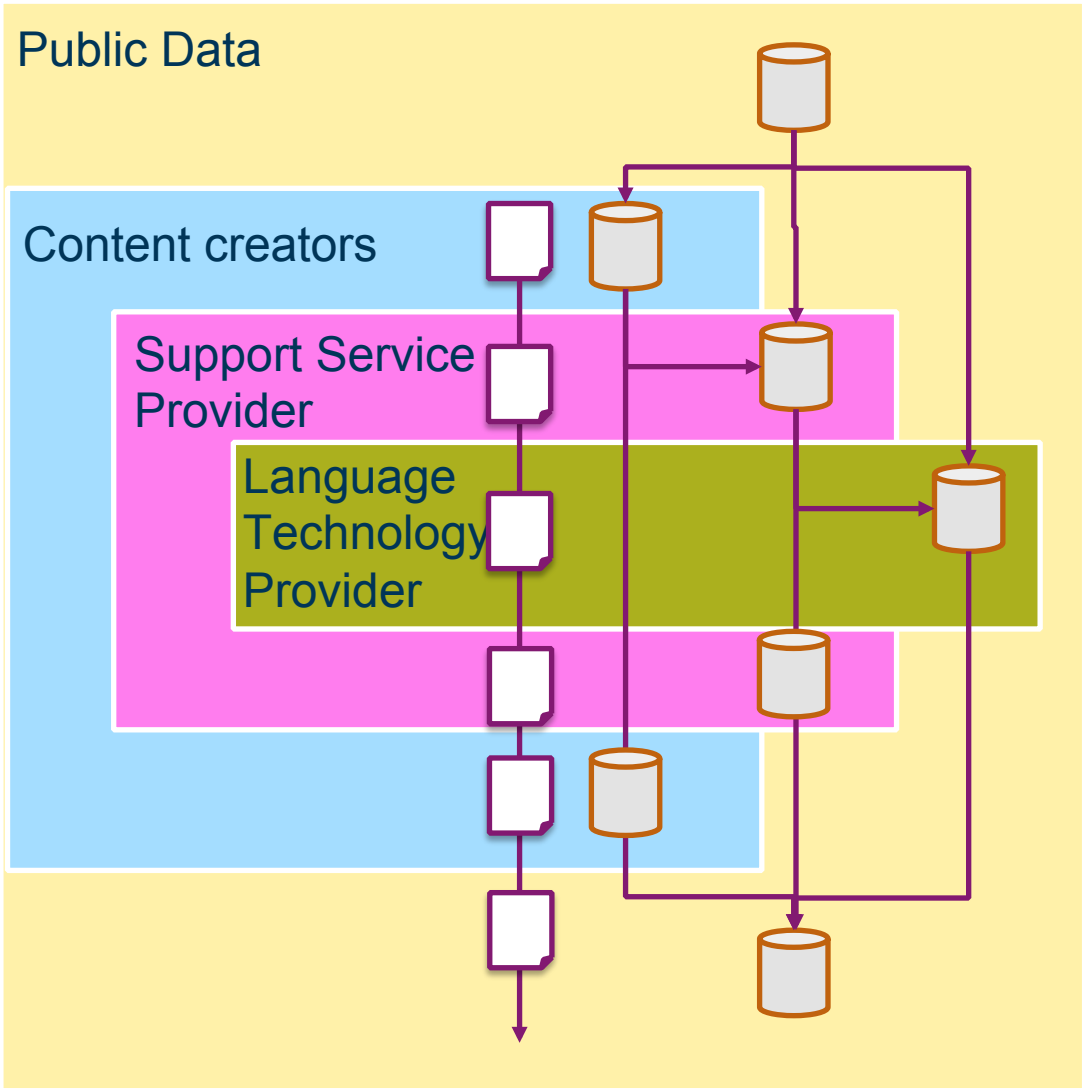
Dave Lewis, ADAPT Centre - TCD

Data in Intelligent Content Processing

- Language Technology relies on Language Data:
  - Corpora
  - Lexical knowledge
  - Domain knowledge/ontologies
- BUT Data Quality and Relevance are Key

- Open Data on the Web has enabled of Massively Multilingual Language Data Aggregators
  - E.g. BabelNet, DBPedia
- BUT how to integrate with domain/proprietary data and maintain shared data quality/relevance

Integrated Content-Data Lifecycles



- Content value chains well established
  - Content creators
  - Support service providers , e.g. translation, analytics
  - Language Technology providers, e.g. MT, text analytics
- Challenge: How to integrate data and content value chains?

Pipeline of Relevant Standards

- Established
  - Provenance Vocabulary
  - Data Catalogue
  - JSON-Linked Data
  - ITS2.0 Vocabulary
- In progress
  - CSV of the Web: tables and JSON meta-data
  - Open Annotation
- Emerging:
  - Ontolex - common lexical-semantic vocabulary
  - Open Data Rights Language
  - Provenance Plan – data flow models

Interoperability Targets

- Open APIs – Basic lexical services
  - Spell checkers
  - Part of speech taggers
  - Sentence splitters
  - Stemmers
- Open Text Analysis Services – lexical-semantic
  - Automated Term Extraction
  - Named Entity Recognition
  - Work Sense Disambiguation
- LT Service Chains must handle overlapping annotations
- Common lexicons vs. domain terminology/ontologies
  - Needs mechanisms to grow, merge and maintain lexical semantic graph across value chain

- THANK YOU!