META FORUM 2015

# A Summary of Research Activities: Technologies – Demands – Gaps – Roadmaps

**Dave Lewis**

**ADAPT Centre, Ireland**
Dave.Lewis@scss.tcd.ie

**META-FORUM 2015: Technologies for the Multilingual Digital Single Market**

Riga, Latvia, April 27, 2015

META-NET has received funding from the EU's Horizon 2020 research and innovation programme through the contract CRACKER (grant agreement no.: 645357). Formerly co-funded by FP7 and ICT PSP through the contracts T4ME (grant agreement no.: 249119), CESAR (grant agreement no.: 271022), METANET4U (grant agreement no.: 270893) and META-NORD (grant agreement no.: 270899).

# Outline:
# Text Analytics and Data

- Overview on Contributing Projects
- Current Technologies, Trends, Directions
- Current Technology Solutions
- Current and Future Demands
- Future Trends, Directions, Technologies
- Language Technology Services, Platforms, Infrastructures

# Current Technologies, Trends, Directions

- Text Analytics relies on Data:
  - Corpora
  - Lexical knowledge
  - Domain knowledge/ontologies
  - Quality and Relevance are Key
- Open Data on the Web has enabled of Massively Multilingual Language Data Aggregators
  - E.g. BabelNet, DBPedia
- Service Composition is key:
  - Established composition platforms and libraries for reuse: UIMA, GATE
  - Slow deployment onto cloud: PANACEA, cloudgate
  - Not yet integrated with Open language data: Datahub, Linghub, NIF
  - Many Data and Service Interoperability challenges remain

# Contributing Projects

META NET

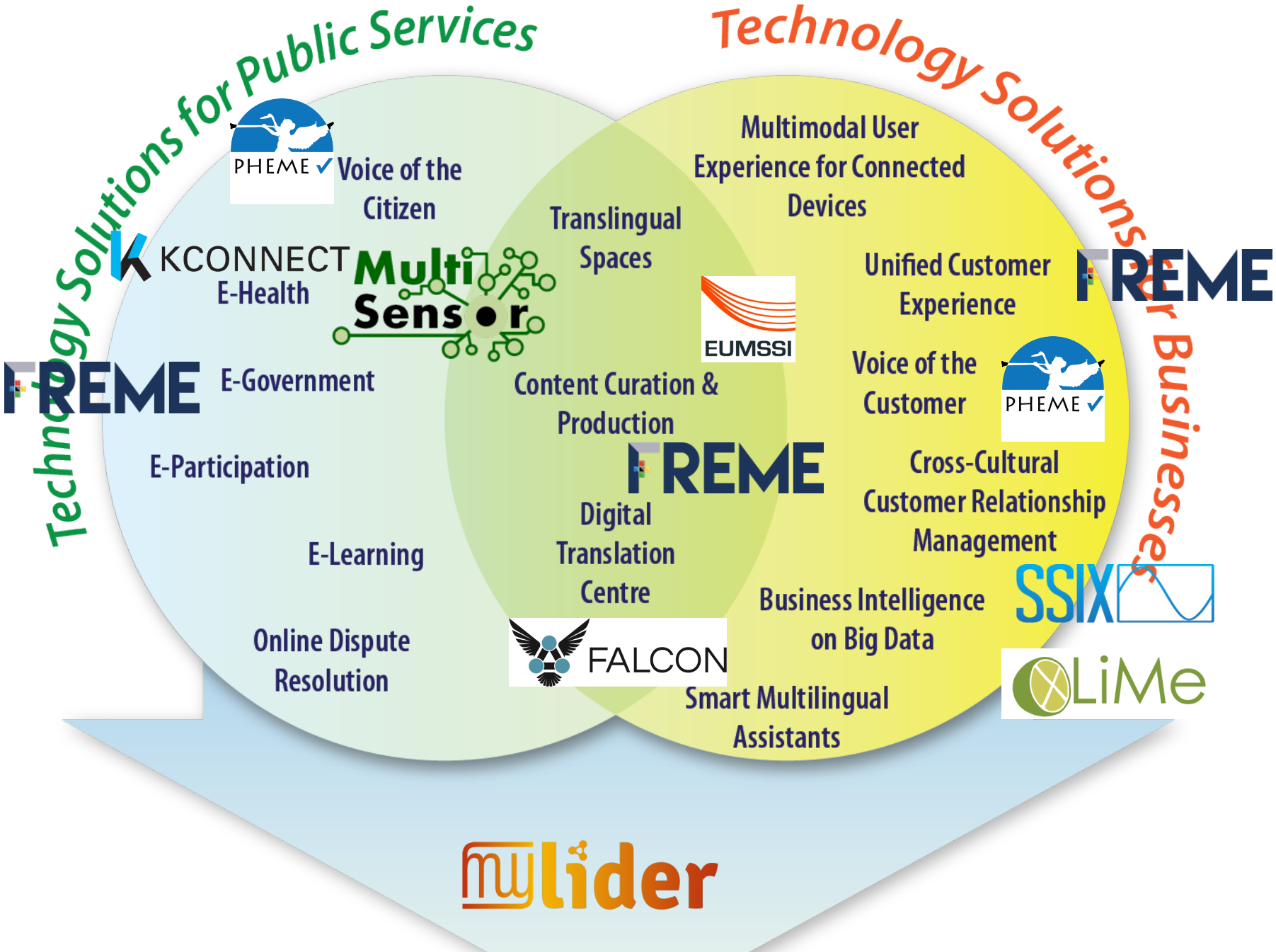









- Kalina Bontcheva (Sheffield) http://www.pheme.eu/
  - Identifying Rumerous Memes
- Paul Buitelaar (INSIGHT) MixedEmotions
  - Social semantic sentiment analytics
- Brian Davis (INSIGHT) http://www.sixx-project.eu
  - Social/sentiment news media analysis
- Asun Gomez Perez (UPM) http://www.lider-project.eu/
  - Linguistic Linked Data
- Allan Hanbury (Vienna UT) http://Kconnect.eu
  - Semantic annotation, search and MT for medical records
- Dave Lewis (ADAPT-TCD) http://www.falcon-project.eu/
  - Linked Data for Localisation
- Maite Melero (UPF) http://www.eumssi.eu/
  - Multimodal media stream analysis
- Achim Rettinger (KIT) http://xlime.eu/
  - Semantic analysis of events/sentiment/opinion over multiple media channels
- Felix Saski (DFKI) http://www.freme-project.eu/
  - Multilingual and semantic enrichment of digital content with data
- Stefanos Vrochidis (ITI-CERTH) http://multisensorproject.eu/
  - Multilingual sentiment, social, spacio-temporal Media Analysis


Technology Solutions for Public Services
Technology Solutions for Businesses
PHEME
Voice of the Citizen
KCONNECT
E-Health
MultiSensor
FREME
E-Government
E-Participation
E-Learning
Online Dispute Resolution
Translingual Spaces
EUMSSI
Content Curation & Production
FREME
Digital Translation Centre
FALCON
Multimodal User Experience for Connected Devices
Unified Customer Experience
FREME
Voice of the Customer
PHEME
Cross-Cultural Customer Relationship Management
SSIX
Business Intelligence on Big Data
LiMe
Smart Multilingual Assistants
mylider

# Future Trends, Directions, Technologies

- Integrate power of natural language processing and semantics
  - Content-semantic interlinks
  - e.g. Semantic search, entity linking
- Linguistic Linked Data
  - Decentralised publishing and discover of language resources
  - Melding linguistic and domain knowledge, e.g. DBPedia, BabelNet
- Highly contextualised multimodal media analytics
  - Reacts to physical, social, temporal context
  - e.g. Search without Querying, sentiment analysis

# Current and Future Demands

- Massive Demand for Data and Content Analytics
  - Advent of commercial data scientist & social media analyst
- Make LT accessible to SME and Public Sector
  - Lower Technical Barriers
  - LT engineering for programmers and app developers
- Integrate Data Management and Content Management
  - Lifecycles and tool chains
  - LT use is Measurable, Predictable, Repeatable
- Text Analytics in Multimedia, Multimodal applications
  - Analytics across channels and media types
- Scale analytics across languages with predictable costs

# Language Technology Services, Platforms, Infrastructures: **Technical**

- Lower the technical entry cost for use of Language Technology & Linguistic Linked Data
  - Enable innovation at web/app speeds and skills
  - Web APIs, cloud and mobile implementations
  - Multiple data bindings: JSON, XML, CSV …
  - Tools for QA, cleansing & maintenance of data
- Best practice, language data management guidelines
  - Targeted to app developers, software engineers, linguists, data curators
  - Public Sector & SME friendly, localised

# Language Technology Services, Platforms, Infrastructures: **Services**

- Open Services impossible without Open Data
- Open Services with open data
  - Build open source communities not just code
  - Public services as innovation platform:
    - Spell checkers, part of speech taggers, sentence splitters, stemmers
    - Limit so as to not impede innovation
  - Common lexicons vs. domain terminology
- Proprietary Services with open data
  - Open data formats, application specific profiles with industry bodies
  - Objective techniques for comparing data quality and service effectiveness, e.g. user A/B testing
  - Enable pipelines & mash-ups between vendors

# Language Technology Services, Platforms, Infrastructures: **EU Policy and Investment**

- EC level:
  - Develop Royalty-free Guidelines, Standards, Best Practice with existing international bodies: W3C, OASIS, WikiCommons
  - Profile vocabulary use for language license, catalogue, provenance meta-data
  - Free data Conformance Test services and live uptake Observatories
  - Indexing and search services
  - Networking/sharing public sector or civil society initiatives

- EU Policy – implement at European and national levels:
  - Every publicly funded Translation published as aligned bi-text web data, every Definition as open onto-lexical data
  - Every publicly funded scientific publication publishes experimental data interlinked to paper

# Priority Research Themes

- Hybrid Statistical and Knowledge-based LT
  - Interlinking corpora and lexical-conceptual resources
  - Machine learning over interlinked graph of corpora, annotations, lexical knowledge and domain ontologies
- Stream Analytics & Stream Resources
  - Analytics over multi-modal, multi-channel, multi-preference content streams
  - From centralised data repositories to live curation/correction/ contextualision of feeds
  - Graph change feeds, dynamic graph mappings and overlays

# Priority Research Themes

- Enabling Multilingual Data Value Chain
  - Instrument to Measure the Value of linguistic data in an application
  - Visualise data value in Application Contexts
  - Predictability and Repeatability of value of language data in efficacy of media analytics and adaptivity
- Manage the ownership, rights and rewards
  - Protect & Pool niche knowledge
  - Encourage social production-ownership of linguistic data
  - Privacy by Design for Social Media data resource
  - Tools for Governance and Transparency

# Priority Research Theme

- Scaling to the Digital Single Market
  - Reduce marginal cost of adding language or domain
  - Porting of linguistic knowledge across languages
  - Managing quality of data and interlinks at scale

- Integrating Text & Media
  - Challenging data scalability & High Variety of media formats
  - Language analytics in context:
    - dialogue, thread, meme, pheme, stream
    - event/time, social, physical, domain task
    - emotion, affect, intent, para-linguistic

# Q/A

**Thank you.**

**office@meta-net.eu**

**http://www.meta-net.eu**
**http://www.facebook.com/META.Alliance**