FALCON

The Localization Web:

From L10n workflows to Linked Data

Dave Lewis
CNGL at Trinity College Dublin

CNGL CENTRE FOR GLOBAL INTELLIGENT CONTENT
XTM International
Interverbum Tech
easyling 1-click website translation

Problem

FALCON

- Multilingual web pages could offer an important language resource,
  - e.g. as parallel text for machine translation engine or multilingual term extraction
- Difficult to leverage, HTML is a publication format, it hides valuable translation info:
  - Translated sentence alignment
  - Term meta-data
  - Translation provenance: was it machine translated, transcreated, quality checked?
- Barrier to leverage by industry's long tail of SME LSPs and clients

The Localization Web

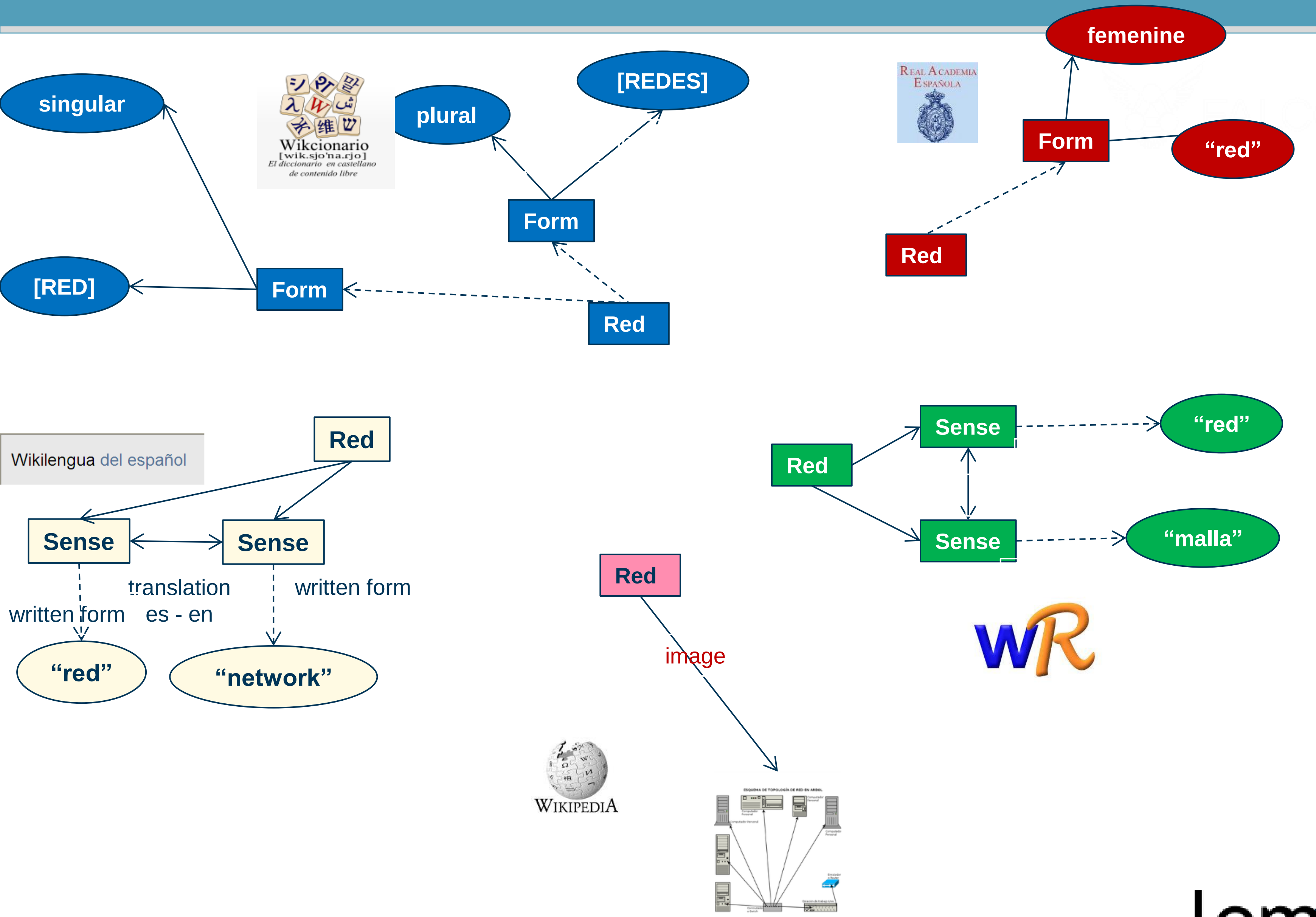
FALCON

- W3C Semantic Web standards allow data to be published on Web
  - Fine-grained URI-based inter-linking
  - Extensible meta-data
  - Standard Query APIs
- Enables a Localization Web
  - Terms and translations become linkable resources
  - Meta-data from L10n workflows adds value
  - Leverage in training Machine Translation and Text Analytics

The Localization Web = Decentralised Annotated Global Translation Memory and Term Base

LD allows linguistic data integration

FALCON



Wikcionario
[wik.sjo'na.rjo]
El diccionario en castellano de contenido libre

singular

plural

[REDES]

Form

Form

[RED]

Red

REAL ACADEMIA ESPAÑOLA

femenine

Form

“red”

Red

Wikilengua del español

Red

Sense

Sense

translation
es - en

written form

written form

“red”

“network”

Red

Sense

Sense

“red”

“malla”

Red

image

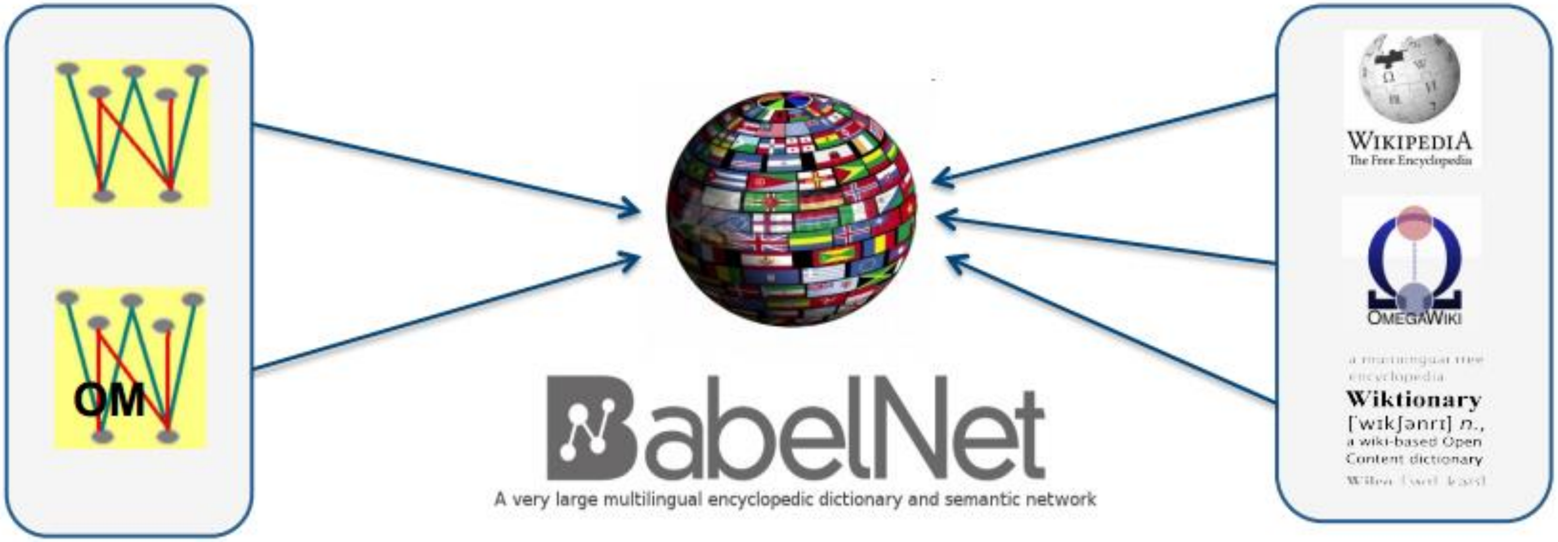
WIKIPEDIA

lemon

BabelNet

Traditional lexical resources

Collaborative lexical resources



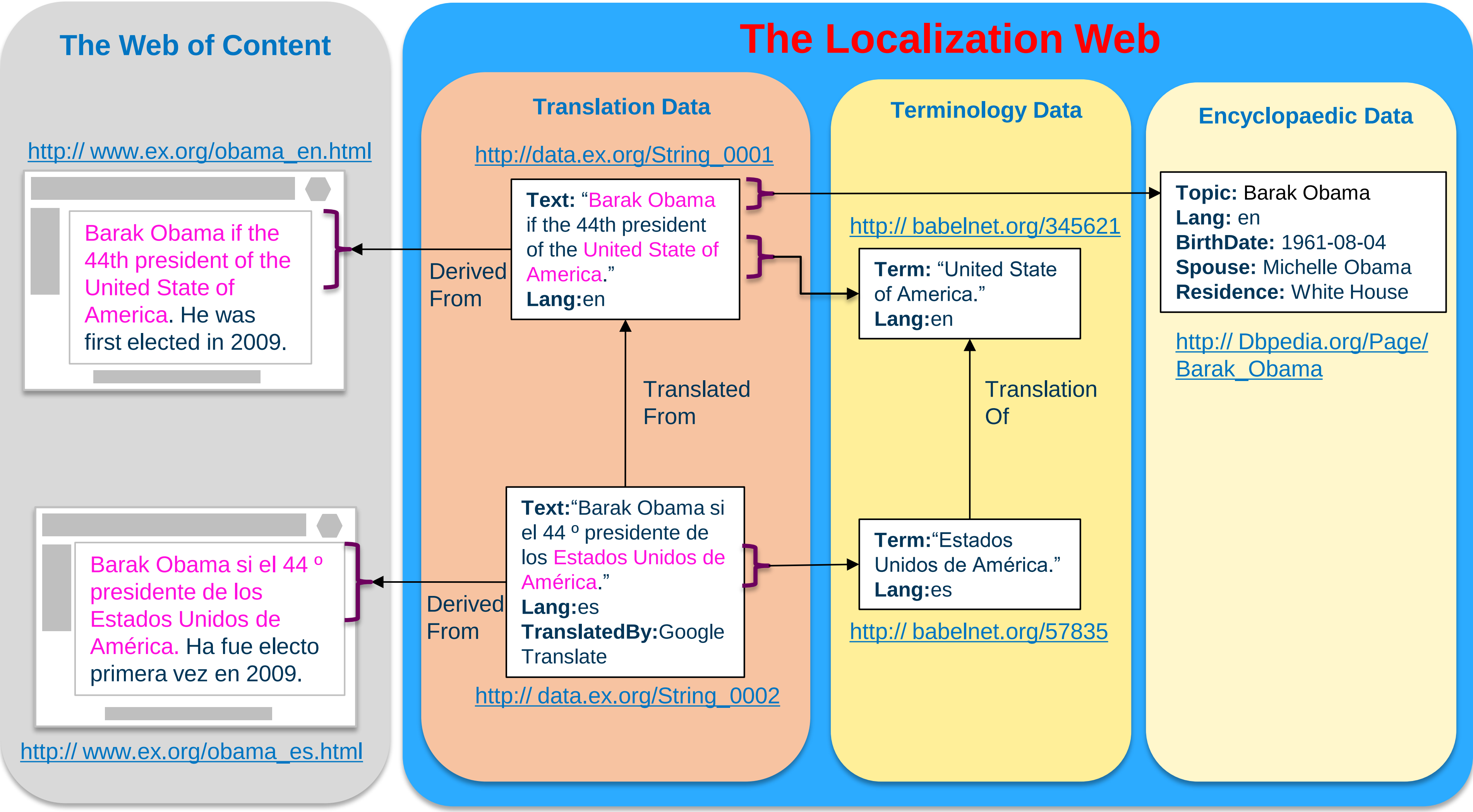
Traditional lexical resources:

- fully-structured
- manually curated by experts
- available for a few languages
- difficult to maintain and update

Collaborative lexical resources:

- semi-structured
- collaboratively built by the crowd
- highly multilingual
- up-to-date

Words as Resources on the Web



The Web of Content

http:// www.ex.org/obama_en.html

Barak Obama if the 44th president of the United State of America. He was first elected in 2009.

Barak Obama si el 44 º presidente de los Estados Unidos de América. Ha fue electo primera vez en 2009.

http:// www.ex.org/obama_es.html

The Localization Web

Translation Data

http://data.ex.org/String_0001

Text: “Barak Obama if the 44th president of the United State of America.”
Lang:en

Derived From

Translated From

Text:“Barak Obama si el 44 º presidente de los Estados Unidos de América.”
Lang:es
TranslatedBy: Google Translate

Derived From

http:// data.ex.org/String_0002

Terminology Data

http:// babelnet.org/345621

Term: “United State of America.”
Lang:en

Translation Of

Term:“Estados Unidos de América.”
Lang:es

http:// babelnet.org/57835

Encyclopaedic Data

Topic: Barak Obama
Lang: en
BirthDate: 1961-08-04
Spouse: Michelle Obama
Residence: White House

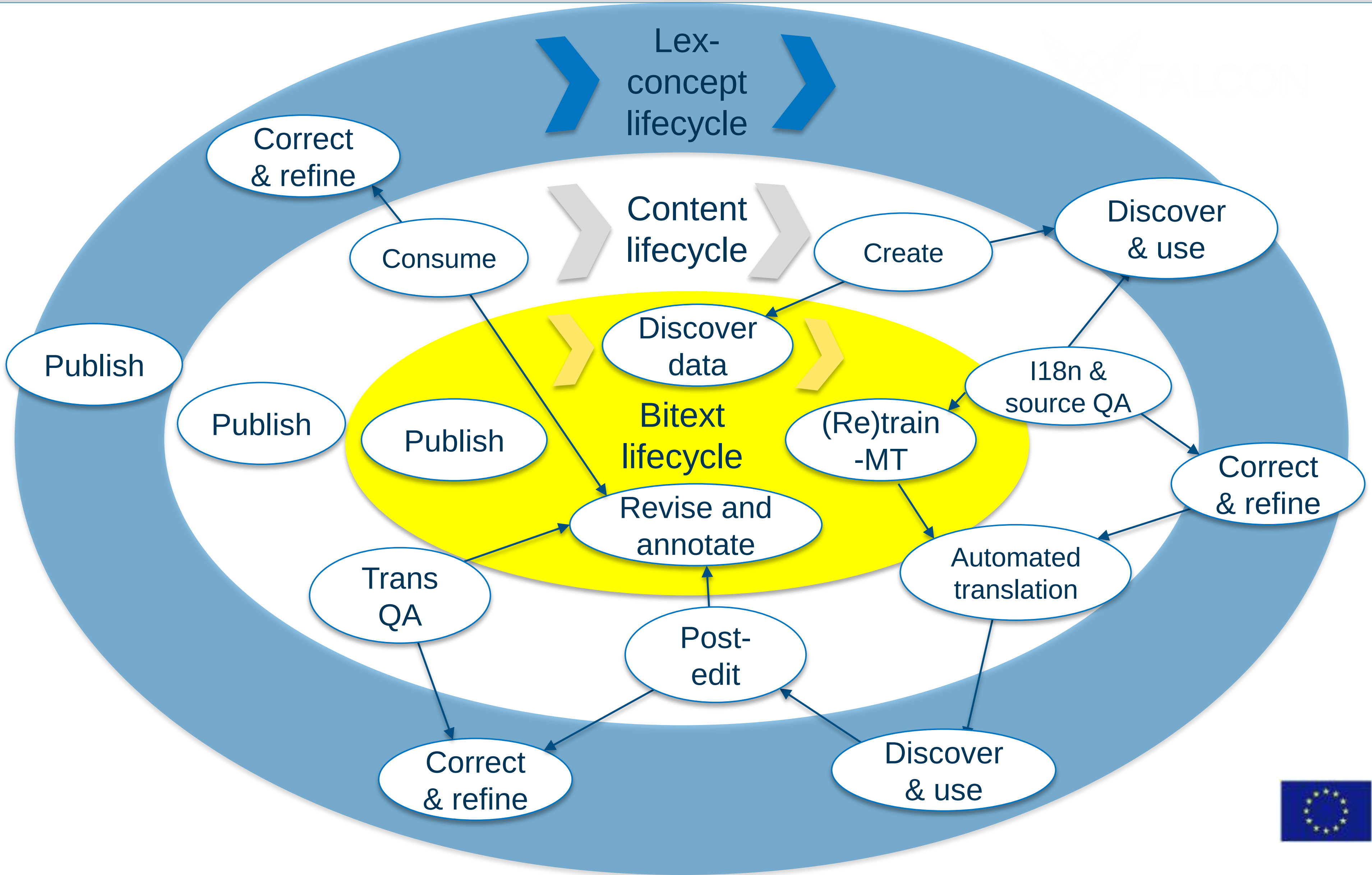
http:// Dbpedia.org/Page/Barak_Obama

Use Cases

- Source Internationalisation
  - Term disambiguation
  - Term extraction with translation discovery
  - Auto-tag named entities with encyclopaedic reference for authors and translators
- Machine Translation
  - Consistent machine translation of terms
  - Pooling and discovery of parallel text for training
- Translation and Post-editing
  - Term definitions from open encyclopaedic data
  - Concordancing over a global TM

Data Management Lifecycles

FALCON



Realisable Benefits?

- How can Language Resource Publishers can audit links to and use of resources & track ROI?
- Can Tool Vendors and Integrators expand markets with more open asset management offerings?
- How can SME LSPs gain resource sharing and pooling opportunities and avoid lock-in?
- Can LSPs and clients can use Active Curation to quickly train domain specific SMT and text analytics components?

Data Management Needs?

- Assert ownership and attribution, licensing, access control?
- Persistent URLs?
- Open royalty free standards?
- Indexing is key, federated vs aggregated data?
- Third party submission of errors, QA, corrections? Publish action on submissions?

Data Access Needs?

- Query bitext on: languages, terms, MT engine, MT confidence, QA, translator qualfications, postedit characteristics?
- Query term base on: language, domains, context, semantic relations, provenance of lexical/conceptual data?
- Web API:
  - HTTP content negotiation (unicode extensions for translation?),
  - Format: RDF, TMX, TBX, RDF, JSON-LD,
  - SPARQL?

More Information

- Contact: dave.lewis@cs.tcd.ie
- http://www.falcon-project.eu

- See also:
  - Linked Data for Language Technology (LD4LT) W3C Community Group
  - http://www.w3.org/community/ld4lt/